DISCRIMINATIVE RELEVANCE FEEDBACK WITH VIRTUAL TEXTUAL REPRESENTATION FOR EFFICIENT IMAGE RETRIEVAL

Suman Karthik[1], C.V. Jawahar[2]

[1] [2]Center for Visual Information Technology, International Institute of Information Technology Hyderabad
Gachibowli, Hyderabad - 500032, India
sumankarthik@gmail.com, jawahar@iiit.ac.in

Keywords: Region Based indexing, Virtual Text, Bag of Words, Image Retrieval

Abstract

The state of the art in contemporary visual object categorization and classification is dominated by "Bag Of Words" approaches. These use either discriminative or generative learning models to learn the object or scene model. In this paper, we propose a novel "Bag of words" approach for content based image retrieval. Images are converted to virtual text documents and a new relevance feedback algorithm is applied on these documents. We explain how our approach is fundamentally different to existing ones and why it is ideally suited for CBIR. We also propose a new hybrid relevance feedback learning model. This merges the best of generative and discriminative approaches to achieve a robust and discriminative visual words based description of a visual concept. Our learning model and "Bag Of Words" approach achieve a balance between good classification and efficient image retrieval.

1 Introduction

In the early days of Content Based Image Retrieval (CBIR) global feature based image retrieval was prolific. These schemes used primitive features of color, shape and texture over the entire image to retrieve relevant images. The shortcomings of such schemes is mentioned in detail in [1]. Later, spatial layout based schemes sampled images in finer detail by dividing them into many small parts usually equal sized and extracting the local features from each part. This evolved into the paradigm of Region Based Image Retrieval [3, 4, 12]. In this general framework, the image is segmented into different homogeneous regions based on either colour, texture, shape or all three of them. These schemes range from segmenting the image into objects to segmenting them into homogeneous color patches. These schemes model the way in which humans percieve visual content better and there by obtaining better performance. However, accurate object segmentation in general is very costly in terms of computational resources. On the other hand, inaccurate segmentation leads to drop in precision of retrieval. Research along this direction came into its own with pioneering work done by Carson *et al*. in their blobworld system [3]. Since then many improvements have been suggested to the general approach of region based image retrieval, the most notable of which was the work done by Wang *et al*. [4, 12].

In recent years, great strides haven been taken forward in the field of visual object categorisation and classification both in images [10] and videos [9]. Many of these approaches use either generative, heirarchical [11], discriminative, or hybrid [5], learning models to learn and classify object categories. This success has been in no small part due to the excellent array of local scale and affine invariant detectors and robust descriptors [7, 8] calculated from the detected points. Their ability to caputre the visual essence of an object and their robustness to real-world imaging situations, makes them the corner stone in the development of efficient and robust object recognition and classification schemes. These local features form the visual words in the bag of words models.

The Bag of Words approach is borrowed from text document categorization and classification. It is generally observed that a collection of commonly occuring phrases can loosely describe a concept or category of documents. This, when adapted to object categorization and classification means that each word is represented by a local descriptor. A collection of these local descriptors are used to describe, discriminate, or categorize, a concept or an image.

This same approach however cannot be directly adopted to Content Based Image Retrieval(CBIR). Though the problems being tackled might look the same, they are infact very different in nature. CBIR often needs to retrieve images belonging to similar concept, but with very little visual similarity. This broader definition of CBIR renders all highly discriminative and highly representative local features obsolete. In CBIR, one must try and achieve concept classification with features that should be primitive enough to adapt to any visual conditions. Hence, CBIR requires robust learning models to improve retrieval performance built on primitive visual features. In this paper, we formulate CBIR as a text document retrieval problem to enhance both the efficiency and the learning capabilities of the scheme.

We propose a novel general representation where images are

treated as documents, and segments are treated as keywords. The virtual textual representation transforms the CBIR problem into a modified text retriveal problem, thereby allowing us to use the wealth of knowledge to tackle the general problems in CBIR (Section 2). We demonstrate the use, practicality and performance of our virutal textual representation scheme with an example implimentation and a pictorial example. Using this representation, we develop a discriminative relevance feedback scheme creating a unique blend to improve both performance and flexibility. The proposed relevance feedback scheme, tries to find the discriminative regions instead of the salient regions to improve the retrieval(Section 3). These regions are discovered in a way that can aid long term learning and at the same time refine the results at each iteration. We validate our scheme under different conditions through a series of experiments(Section 4). We also show that our scheme can be extended to achieve better performance without trading it for flexibility.

2 Virtual Textual Description

A sunset described visually in terms of color by a human would be something as follows.

sunset $\rightarrow$ (Orangish *or* Reddish) Hue on Top AND (Yellow *or* Bright Yellow) Hue in the middle

Human beings tend to describe visual content as a group of visually coherent regions. Hence we can see that the sky is expressed as orangish or reddish hued region on top. Such a general description of a sunset allows for a lot of variation as does the human recognition of a generic sunset. The concept of 'Sunset' is, by definition, visually and conceptually broad and inexact in nature. This broad description allows us and the scheme to accomodate other visually different concepts like clouds and buildings in the sunset image.

An image can be described and distinguished as a collection of regions or segments in order to better handle the content. Here the image becomes a collection of discrete visual concepts that are put together to form one visually coherent concept. This is like a bunch of words put together to form a coherent essay, document, or description. We hence draw the parallels between the logical compactness of words and segments in images and documents. For example we see that for a concept sunset Orangish, Reddish, Yellow, Bright Yellow are keywords in textual form. This is carried on into the image domain, where images are modeled as text documents and segments are keywords of these documents. Such a modeling tries to mimic human visual interaction or description rather than human visual perception. Hence visual concepts can be communicated effectively between the user and the system.

In our scheme, an image is treated as a visual document akin to a text document and the major or the important segments of the image are treated as keywords in the text document. Once the image is segmented each segment is visually described in the form of a word where the word is a 6 character string instead of linguistic representation like "Orange" or "Blue". This word is the result of binning visual features of the image and applying a linear transformation to obtain a 6 character string in the text domain. This six character string is called a "keyword" and each image is called a "Document". The nature of these Keywords is such that they are inherently broad or inexact representations of their respective segments unlike numerical representations. In our scheme, the distance between two documents cannot be calculated by cosine distance as in document retrieval. This is because the keywords themselves have a distance between them which incorporate more fuzziness into the scheme and as a consequence robustness. We use hamming distance to calculate distance between two keywords and hence two segments. Consequently least cumulative hamming distance between two images produced by any configuration is used as the "Inter Document" or "Inter Image" distance.

A representation of an image as a document and segments as keywords, allows us to pose the CBIR problem as a special "Text Document Retrieval" problem. Such a transformation has the promise to improve the ability to index and retrieve images based on content using accumulated knowledge and practices in the text document retrieval domain. Existing proprietary or open source database systems can be used to store and index the images and also to efficiently retrieve these images. This would not be possible using the conventional feature based representation and spatial databases would have to evolve. Our representation can become translation and transformation independent as and when required automatically by dropping the importance associated with positions of the segments. Our scheme can also handle occlusion as the segments are independently modelled, and occlusion of one or more of the segments will be handeled gracefully.

2.1 An Example Implementation

The image is initially mapped into an appropriate color space where the human visual perception is much more concordant. This image is then quantized into a discerete number of uniform bins in the feature space. The image is then segmented based on the color and spatial constaints. The segmentation algorithm is a heuristic algorithm designed to be much more robust and handle occlusion or collection of similar objects. The segmentation is very efficient when compared to other contemporary implementations[4, 12] of region based retrieval. It can afford this efficiency because of the concept refinement features built in to the scheme through relevance feedback that make up for the loss of segmentation accuracy.

Once segmented, each segment is treated as a visual word. This

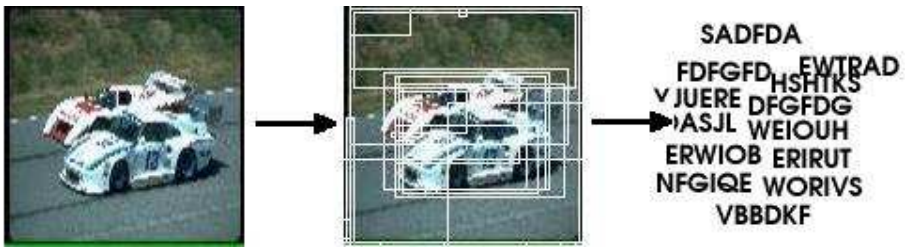


Figure 1: An example of an image being converted into virtual textual representation. First the image is segmented into different parts or visual words, then these parts are transformed into words by quantizing the individual colour, texture and shape features within each visual word. Finally we have a virtual textual representation of the image

visual word is converted into text by a linear trasnformation as shown in the Figure1 above.

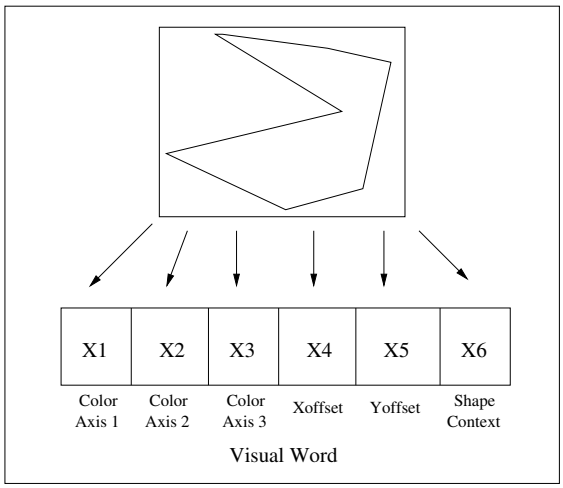


Figure 2: The above figure demonstrates how a visual word is converted into a text or symbol representation in the example implementation, here X1, X2, X3 are the symbols assigned to quantized bins in the colorspace. X4 and X5 are the quantized x and y offset of the segment from a reference and X6 is the shape context of that particular image.

When an exemplar image is given as a query, its representation (collection of all the keywords) Q is extracted by the feature extraction module, where Q_i is the ith keyword in the document. Every other image document K_j is compared with Q to obtain a similarity score S_j for images documents Q and Kj.

$$S_j = \prod_{i=1}^{n} max(H_{k=1}^{m}(Q_i, K_{jk})) \quad (1)$$

$$H = (6 - hammingdistance + 1) \quad (2)$$

Where n is the number of keywords in Q and m are the number of keywords in K_j. Once we get all the S_j we have.

$$\left(\; S_1 S_2 S_3 \ldots S_{m-1} S_m \; \right) \quad (3)$$

We then sort the S_j and take the top N images or documents as the most relevant. This scheme is also very efficient as the problem has been modeled into a partial string matching problem, where earlier floating point calculations were heavily used. Now the calculations can be made with simple bit operations instead of costly floating point operations. The above described linear transformation is but an example of a way in which an image can be trasnformed into a symbolic or textual representation. This however might not be suitable for all situations, for example situations where there are really dense cluster seperated by sparse spaces in the feature space. Hence different situations would require different quantization schemes but the general framework of the scehme will remain consistent. Usually in a normal region based image retrieval, if 50 to 70 segments are produced and each segment is described by 6 to 7 floating point numbers as features. In our case we use 6 to 7 symbols to represent each feature vector, or a 6 character string. Already space efficiency is achieved by our representation. Further, each floating point distance computation (minkowski) involves several complex arithemetic operations like square root, cube root, addition and subtraction. This makes floating point based region based image retirieval 50^2 to 70^2 times more inefficient when compared to global feature based methods. Our method on the other hand uses bit operations and text indexing to achieve almost quasi linear execution performance, making it atleast 10 times more efficient than the traditional schemes.

Concept	Images	DRF	Bayesian
Bus	30	82	58
Car	34	98	62
Flower	30	63	42
Rocks	29	60	29
Sunset	35	92	56
Surfers	28	56	31
Train	30	74	54

Table 1: The above table contains 4 columns for dataset D1 as follows. Column 1 contains the class of images. Column 2 contains the number of images from each class Column 3 contains the precision of Discriminative relevance feedback(DRF) Column 4 contains the precision of a simple bayesian relevance feedback approach(Bayesian)

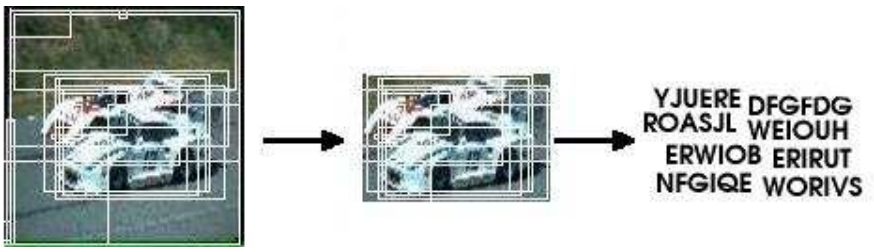


Figure 3: The different words or image patches that make up the car are further refined during discriminative relevance feedback and a only the most discriminating words are retained. This impoves both the classification performance and the efficiency of the scheme.

3 Discriminative Relevance Feedback

Recent years have seen the development of many relevance feedback strategies for region based image retrieval as in the work done by Jing *et al*. [6]. But most of the existing systems still use relevance feedback techniques built for global feature based image retrieval. Other region based relevance feedback algorithms make use of region weighting to achieve retrieval. Such techniques do not effectively distinguish a class of images in the presence of other classes in the database. Rather they tend to cluster images based on the nature of the relevant class which may lead to accidental biases toward unimportant features or regions. At the same time not much work or attention has been given to the efficiency and indexing of region based image retrieval schemes. Our relevance feedback scheme differs from contemporary relevance feedback schemes. Most of the schemes try to either obtain a region weighting or try to extract the regions of these images based on which regions are most dominant in the relevant images. Such schemes have a tendency to become biased toward features that do not actually represent the concept. Other schemes finding the most salient regions in an image because which can also lead to similar bias. For example a couple of "Red Buses" will lead the system to deduce that the regions with red are the important regions for the concept "Bus" which is clearly not the case.

In our relevance feedback scheme we obtain the most discriminative regions or keywords instead of the important keywords of a particular class of images. Given a set of retrieved images R and once the user marks all the relevant images P and the rest are the set of irrelevant images N we calculate the most discriminative keywords. This is done by defining a "Segment To Image" or "Keyowrd To Document" distance D_{si} which represents how close a segment or keywords is to an image. If SEG is the set of all the segments of P, then a pseudo-image of top num whose cumulative distance to images in P is the least and the cumulative distance to images in N is the highest. This is quantitatively represented by a discriminability measure for each keyword in P calculated as discussed below. Hence we make a new pseudo-image with the most discriminative keywords of image class represented by R, allowing us to pick the representative segments dependent on the other classes in the database. This is done over many iterations.

As the relevance feedback scheme used tries to pick what makes each class unique, this uniqueness can be easily captured to aid in learning the concepts in the long term. As the scheme is flexible, with slight modifications anything from spatial constraints to optimal segment grouping can be incorporated to achieve better results. Such a scheme will aid in distinguishing visually similar looking concepts. Once these keywords are obtained we make a pseudo-image or document out of the most discriminative keywords. This pseudo-document is refined over further relevance feedback iterations. Hence in the end we have keywords or segments that are able to represent very specifically the concept they represent.

3.1 Algorithm

1. Obtain query image Q.
2. Obtain the image document (Collection of Keywords).
3. Image set R is retrieved from the database by the nearest neighbour retrieval algorithm.
4. Obtain feedback from user on R as P set of relevant image documents and N set of irrelevant image documents.
5. Calculate the most discriminative keywords from P and N
   - Calculate the Relevance score r_p among P for each keyword in P.
   - Calculate the Relevance score r_n among N for each keyword in P.
   - Obtain discriminative score d_r for all the keywords in P as $\frac{r_p}{r_n}$.
   - Sort the keywords in descending order of discriminative score d_s.
6. Pick top num keywords from the set of keywords such that all of them are mutually dissimilar by a minimum hamming distance of x.
7. Collect these num keywords and construct a new pseudo image document and loop to step 2 until the user quits.

In the above algorithm we can see that only the keywords from P are used to estimate the new image or the pseudo image document of the concept at hand. Here we try to find the regions or keywords that are exclusive to a particular concept rather than keywords that are important to a particular concept. We also provide a threshold for discriminative capability of two regions or keywords using x as the minimum hamming distance because of the need to eliminate redundant regions and at the same time allowing the pseudo image document to be as expressive as possible. Our algorithm can be termed as a hybrid bag of words approach as we are starting out with a generative model of what a particular concept is, then this model is modified by a discriminative learning model that refines the generative model to achieve discriminability from other concepts in the dataset.

4 Results and Analysis

We tested two methods or algorithms discriminative relevance feedback(DRF) and relevance feedback based on region importance(Bayesian). The methods were tested on two image sets $D1$ with 225 images and 7 categories and $D2$ with 1162 images and 15 categories. All the images in the two databases were taken from the corel image database [2]. $D1$ was used to confirm the methods ability to perform under well defined and visually disparate concepts and $D2$ was used to test the robustness of the schemes under conceptually different categories that are visually very similar.The retrieval set was of size 20 and this was used to calculate precision over a number of iterations.

$$Precision = \frac{NumberofRelevantImagesRetrieved}{SizeofRetrievedSet} \quad (4)$$

Here we find that our method DRF clearly outperforms the Bayesian probability based salient region retrieval method. We also observed that our scheme was able to distinguish beautifully between even hard to distinguish categories like "Surfers" and "Waves" or "Flowers" and "Roses", and this is more prominent when one considers that the only features of significance here are 3 color features. Another important observation is that the DRF's precision fluctuates, Bayesian however shows a stable increase in precision in the majority of the cases. Also as the number of distinct concepts grows DRF tends to browse through a wide variety of these classes based on the discriminability. So DRF requires some iterations to get its bearing in the concept space. The performance of DRF on visually coherent concepts is outstanding. This can be clearly seen in the tables of $D1$ and $D2$ above. In both cases the user critiques on wether the given images are relevant or irrelevant. It was assumed the user critiques are consistent and deterministinc regarding the relevance of an image to a concept.

Concept	Images	DRF	Bayesian
Bus	91	88	63
Car	39	85	54
Flower	74	60	48
Cat	58	22	15
Sunset	135	85	40
Surfers	89	54	28
Train	82	66	52
Skiers	65	13	9
Sailboat	64	34	32
Tools	79	81	66
Waterfall	86	30	27
Wave	74	23	2
Bicycle art	78	54	52
Birds	82	34	26
Roses	101	87	56

Table 2: The above table contains 4 columns for dataset D2 as follows. Column 1 contains the class of images. Column 2 contains the number of images from each class Column 3 contains the precision of Discriminative relevance feedback(DRF) Column 4 contains the precision of a simple bayesian relevance feedback approach(Bayesian)

5 Conclusion

In the paper we have described a Region Based Image Retrieval Framework that suggests a modeling of the CBIR problem as a text retrieval problem. We also propose a relevance feedback problem that works primarily on bringing out the discriminative regions of various concept classes. Our scheme is also very efficient as the segmentation algorithm is primitive and at the same time is fuzzy. We have also established that any shortcomings in the primitive yet fast and inaccurate segmentation can be handled by our robust Relevance feedback algorithm. We have also seen that our scheme is even able to distinguish between different concepts that are more or less visually similar in nature. With further refinements this can be a scalable method with inherent indexing capabilities built in the form of the text based keywords.

References

[1] S. Santini A. Gupta R. Jain A. Smeulders, M. Worring. Content based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[2] CalPhotos. Corel image database. at, http://elib.cs.berkeley.edu/corel/.

[3] Chad Carson, Megan Thomas, Serge Belongie, Joseph M. Hellerstein, and Jitendra Malik. Blobworld: A system

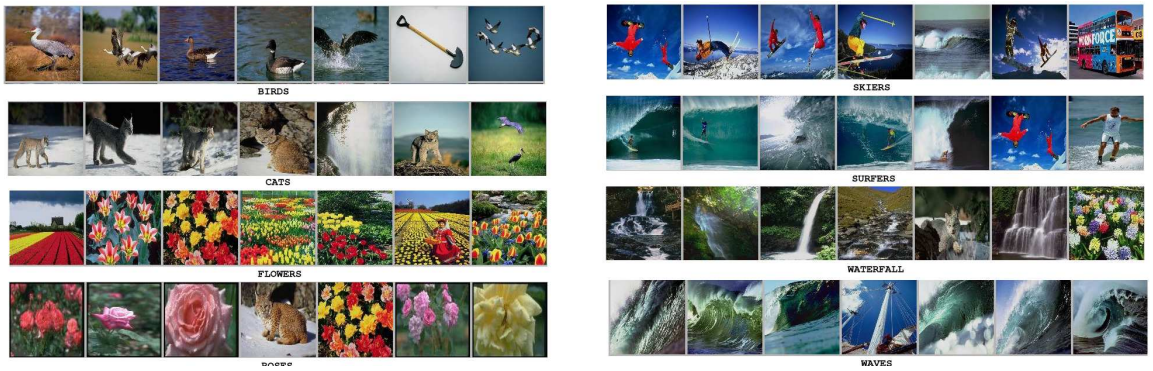
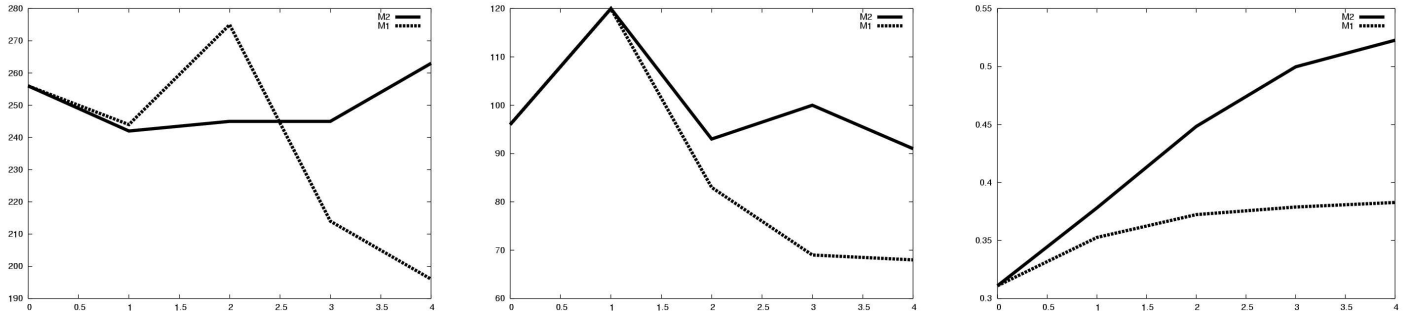


Figure 4: A Small Selection Of Retrieved Results



(a) (b) (c)

Figure 5: (a): Surfers confused as waves with our method and region importance, (b): flowers confused as roses with our method and region importance, (c): Average Precision for our Method and region importance all of them have "Iterations" as X-axis and Images as Y-axis for the (a) and (b) and Precison as Y axis for (c). It can be clearly seen that our method outperforms the general method

for region-based image indexing and retrieval. In *Third International Conference on Visual Information Systems*. Springer, 1999.

[4] Yanping Du and James Ze Wang. A scalable integrated region-based image retrieval system. In *Proceedings Of International Conference of Image Processing (1)*, pages 22–25, 2001.

[5] Mario Fritz, Bastian Leibe, Barbara Caputo, and Bernt Schiele. Integrating representative and discriminative models for object category detection. In *Proceedings of the International Conference on Computer Vision*, pages 1363–1370, 2005.

[6] F. Jing, M. Li, L. Zhang, H. Zhang, and B. Zhang. Learning in region-based image retrieval. In *Jing, F., Li, M., Zhang, L., Zhang, H.J., Zhang, B.: Learning in region-based image retrieval. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer (2003) 198–207*, 2003.

[7] Krystian Mikolajczyk, Bastian Leibe, and Bernt Schiele. Local features for object class recognition. In *Proceedings of the International Conference on Computer Vision*, pages 1792–1799, 2005.

[8] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[9] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, October 2003.

[10] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. Discovering objects and their localization in images. In *Proceedings of the International Conference on Computer Vision*, pages 370–377, 2005.

[11] Erik B. Sudderth, Antonio B. Torralba, William T. Freeman, and Alan S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *Proceedings of the International Conference on Computer Vision*, pages 1331–1338, 2005.

[12] James Ze Wang, Jia Li, and Gio Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.