Content based relevant image retrieval using SVM Clarification

R.B.Gayathri[1], Mathimalar.V[2]
[1, 2] Shrimathi Indira Gandhi College, Trichy, Tamilnadu

Abstract- *The image processing plays a major role in medical and security applications. The image or data transmission and reception can be retrieve back as same as original image. Hence while receiving the image it will be same as original image. To get perfect result this work presents automatically generate a large number of images for a specified object class. A multi-modal approach employing metadata, text and visual features is used to gather many high-quality images from the Web. Candidate images i.e., original image given to the process are obtained by a text-based Web search querying on the object identifier. The given WebPages and its images initially get downloaded. The main task is to remove the irrelevant images present in the subjected image and then re-rank the remainder. First, the images are re-ranked based on the text surrounding the image and metadata features. There is several methods used here to compare re-ranking. By comparing with existing methods the SVM visual classifier is used here to improve the performance of the data or given image. To investigate the sensitivity of cross-validation procedure for noisy training data. The main objective of the overall method is in combining metadata /text and visual features in order to achieve a completely automatic ranking of the images. Examples are given for a selection of , vehicles, animals, and other classes. Our objective in this work is to harvest a large number of images of a particular class automatically, and to achieve this with high precision. Our motivation is to provide training databases so that a new object model can be learned effortlessly.*

Keywords- Image processing; Meta data; Retrieval; Correlation

I. INTRODUCTION

Our objective in this work is to harvest a large number of images of a particular class automatically, and to achieve this with high precision. Our motivation is to provide training databases so that a new object model can be learned effortlessly. Following, we also use Web search to obtain a large pool of images and the Web pages that contain them. The low precision does not allow us to learn a class model from such images using vision alone. The challenge then is how best to combine text, metadata, and visual information in order to achieve the best image re-ranking.

The two main contributions are: First, we show that metadata and text attributes on the Webpage containing the image provide a useful estimate of the probability that the image is in class, and thence, can be used to successfully rank images in the downloaded pool. Second, we show that this probability is sufficient to provide (noisy) training data for a visual classifier, and that this classifier delivers a superior re-ranking to that produced by text alone. It visualizes this two-stage improvement over the initially downloaded images. The class-independent text ranker significantly improves this unranked baseline and is itself improved by quite a margin when the vision-based ranker (trained on the text ranker results) is employed. We compared our proposed discriminative framework (SVM) to unsupervised methods (topic models), concluding that the discriminative approach is better suited for this task, and thus, the focus of this work. Others have used text and images together, however, in a slightly different setting. For example, use ground-truth annotated images as opposed to noisy annotation stemming from Web pages, as in our case. Other work uses text from the Internet, but focuses on identifying a specific class rather than general object classes. We show that our automatic method achieves superior ranking results to those produced by the method and also to that of Google Image Search. The extensions include: a comparison of different text ranking methods, additional visual features (HOG), an investigation of the cross validation to noise in the training data, and a comparison of different topic models (for the visual features).

A. Existing System

- The availability of image databases has proven invaluable for training and testing object class models during the recent surge of interest in object recognition. However, producing such databases containing a large number of images and with high precision is still an arduous manual task.
- Image search engines apparently provide an effortless route, but currently are limited by poor precision of the returned images and restrictions on the total number of Images provided.
- For example, with Google Image Search, the precision is as low as 32 percent on one of the classes tested here

(shark) and averages 39 percent, and downloads are restricted to 1,000 images.

B. Proposed System

- The objective of this work is to automatically generate a large number of images for a specified object class.
- A multimodal approach employing both text, metadata, and visual features is used to gather many high-quality images from the Web.
- Candidate images are obtained by a text-based Web search querying on the object identifier (e.g., the word penguin).
- The task is then to remove irrelevant images and re-rank the remainder.
- First, the images are re-ranked based on the text surrounding the image and metadata features. A number of methods are compared for this re-ranking.
- Second, the top-ranked images are used as (noisy) training data and an SVM visual classifier is learned to improve the ranking further. We investigate the sensitivity of the cross-validation procedure to this noisy training data.

II. SYSTEM MODULE

There are five different types of modules in this project, that are listed in the following,

1. Query Image
2. Download Associate Images
3. Apply Re-ranking Algorithm
4. Filtering Process

A. Query Image

When an image search in search engines, that corresponding images are loaded in that time, meanwhile among them there is a uncategorized images are also spotted. However, producing such databases containing a large number of images and with high precision is still an arduous manual task. Generally Image search engines apparently provide an effortless route. For this type of obtaining images can be filter and arrange. The results of the applicable images are assembled and Our objective in this work is to harvest a large number of images of a particular class automatically, and to achieve this with high precision.

Image clusters for each topic are formed by selecting images where nearby text is top ranked by the topic. A user then partitions the clusters into positive and negative for the class. Second, images and the associated text from these clusters are used as exemplars to train a classifier based on voting on visual (shape, color, and texture) and text features.

B. Download Associate Images

We compare three different approaches to downloading images from the Web.

The first approach, named Web Search, submits the query word to Google Web search and all images that are linked within the returned Web pages are downloaded. Google limits the number of returned Web pages to 1,000, but many of the Web pages contain multiple images, so in this manner, thousands of images are obtained.

The second approach, Image Search, starts from Google image search (rather than Web search). Google image search limits the number of returned images to 1,000, but here, each of the returned images is treated as a “seed”—further images are downloaded from the Webpage where the seed image originated.

The third approach, Google Images includes only the images directly returned by Google image search (a subset of those returned by Image Search). The query can consist of a single word or more specific descriptions such as “penguin animal” or “penguin OR penguins.” Images smaller than 120 _ 120 are discarded. In addition to the images, text surrounding the image HTML tag is downloaded, together with other metadata such as the image filename.

Image Search gives a very low precision (only about 4 percent) and is not used for the harvesting experiments. This low precision is probably due to the fact that Google selects many images from Web gallery pages which contain images of all sorts. Google is able to select the in-class images from those pages, e.g., the ones with the object-class in the filename; however, if we use those Web pages as seeds, the overall precision greatly decreases. Therefore, we only use Web Search and Google Images, which are merged into one data set per object class. Table 2 lists the 18 categories downloaded and the corresponding statistics for in-class and non-class images. The overall precision of the images downloaded for all 18 classes is about 29 percent.

C. Apply Re-ranking Algorithm

Now describe the re-ranking of the returned images based on text and metadata alone. Here, we follow and extend the method proposed by using a set of textual attributes whose presence is a strong indication of the image content.

The goal is to re-rank the retrieved images. Each feature is treated as binary: "True" if it contains the query word (e.g., penguin) and "False" otherwise. To re-rank images for one particular class (e.g., penguin), we do not employ the whole images for that class. Instead, we train the classifier using all available annotations except the class we want to re-rank. This way, we evaluate performance as a completely automatic class independent image ranker, i.e., for any new and unknown class, the images can be re-ranked without ever using labeled ground-truth knowledge (images are divided into three categories: 1.Good, 2.Ok, 3.non-class) of that class.

D. Filtering Process

The text re-ranker performs well, on average, and significantly improves the precision up to quite a high recall level. To re-ranking the filtered images, we applied the text vision system to all images downloaded for one specific class, i.e., the drawings and symbolic images were included.

It is interesting to note that the performance is comparable to the case of filtered images. This means that the learned visual model is strong enough to remove the drawings and symbolic images during the ranking process. Thus, the filtering is only necessary to train the visual classifier and is not required to rank new images,

However, using unfiltered images during training decreases the performance significantly, the main exception here is the airplane class, where training with filtered images is a lot worse than with unfiltered images. In the case of i.e., airplane, the filtering removed 91 good images and the overall precision of the filtered images is quite low, 38.67 percent, which makes the whole process relatively unstable, and therefore can explain the difference.

III. SYSTEM DESCRIPTION

We compare three different approaches to downloading images from the Web. The first approach, named Web Search, submits the query word to Google Web search and all images that are linked within the returned Web pages are downloaded. Google limits the number of returned Web pages to 1,000, but many of the Web pages contain multiple images, so in this manner, thousands of images are obtained. The second approach, Image Search, starts from Google image search (rather than Web search). Google image search limits the number of returned images to 1,000, but here, each of the returned images is treated as a seed"—further images are downloaded from the Webpage where the seed image originated. The third approach, Google Images, includes only the images directly returned by Google image search (a subset of those returned by Image Search). The query can consist of a single word or more specific descriptions such as "penguin animal" or "penguin OR penguins." Images smaller than 120_120 are discarded. In addition to the images, text surrounding the image HTML tag is downloaded, together with other metadata such as the image filename.

Ground-truth annotation.

In a similar manner, images are divided into three categories:

in-class-good. Images that contain one or many class instances in a clearly visible way (without major occlusion, lighting deterioration, or background clutter, and of sufficient size).

in-class-ok. Images that show parts of a class instance, or obfuscated views of the object due to lighting, clutter, occlusion, and the like.

nonclass. Images not belonging to in-class.

The good and ok sets are further divided into two subclasses:

abstract. Images that do not resemble realistic natural objects (e.g., drawings, nonrealistic paintings, comics, casts, or statues).

nonabstract. Images not belonging to the previous class.

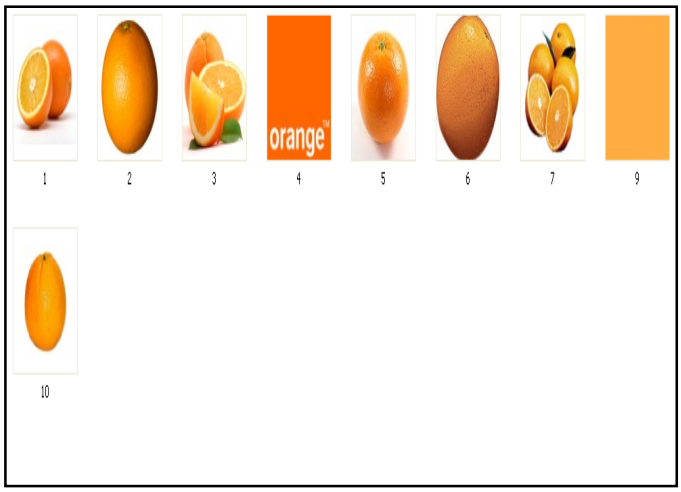

Fig. 1. Relevant and Ir-releveent images

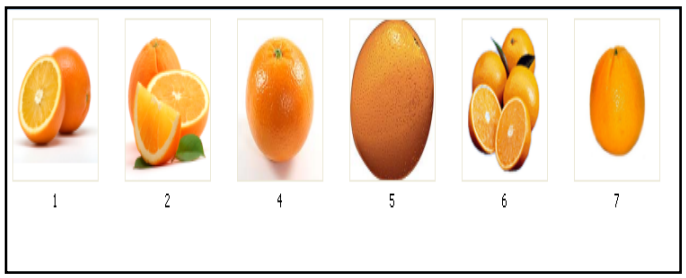

Fig. 2. Relevant and Ir-releveent images

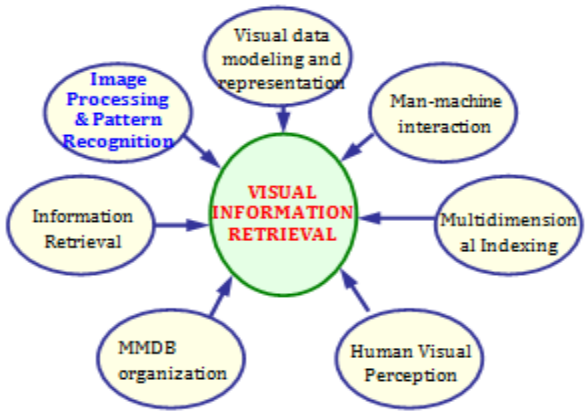

Fig. 3. Visual Information Retrieval blends together many research disciplines

A. Visual information Retrieval

Fig 3 shows the Visual Information Retrieval, (VIR) is a relatively new field of research in Computer Science and Engineering. As in conventional information retrieval, the purpose of a VIR system is to retrieve all the images (or image sequences) that are relevant to a user query while retrieving as few non-relevant images as possible. The emphasis is on the retrieval of information as opposed to the retrieval of data. Similarly to its text-based counterpart a visual information retrieval system must be able to interpret the contents of the documents (images) in a collection and rank them according to a degree of relevance to the user query. The interpretation process involves extracting (semantic) information from the documents (images) and using this information to match the user needs.

Progress in visual information retrieval has been fostered by many research fields, particularly: (text-based) information retrieval, image processing and computer vision, pattern recognition, multimedia database organization, multidimensional indexing, psychological modeling of user behavior, man-machine interaction, among many others.

VIR systems can be classified in two main generations, according to the attributes used to search and retrieve a desired image or video file.

- **First-generation VIR systems:** use query by text, allowing queries such as “all pictures of red Ferraris” or “all images of Van Gogh’s paintings”. They rely strongly on metadata, which can be represented either by alphanumeric strings, keywords, or full scripts.
- **Second-generation (CB)VIR systems:** support query by content, where the notion of content, for still images, includes, in increasing level of complexity: perceptual properties (e.g., color, shape, texture), semantic primitives (abstractions such as objects, roles, and scenes), and subjective attributes such as impressions, emotions and meaning associated to the perceptual properties. Many second-generation systems use content-based techniques as a complementary component, rather than a replacement, of text-based tools.

B. Colour models

The RGB color model is widely used to represent digital images on most computer systems. However, the RGB color model has a major drawback on the similarity measure. This is due to the combination of the color characteristics. Figure 2(a) shows the whole color space of the RGB color model. The lightness and saturation information are implicitly contained in the R, G, and B values. Therefore, two similar colors with different lightness may have a large Euclidean distance in the RGB color space and are regarded as different. This is not consistent with the human perception and will decrease the accuracy of the image retrieval. Some color models, such as HSV and CIE L*u*v*, are proposed to overcome this problem. Their color characteristics are separated into three parts: hue, lightness, and saturation, which make them more consistent with human vision.

In our approach, we choose the HSV color model to represent the color information of an image. The whole color space in the HSV color model is represented by a cylinder, as shown in Figure 2(b). In the HSV color model, the color characteristics are separated into three parts: hue, saturation, and value. Because the total number of colors in the HSV color model is too high, it is necessary to partition the whole HSV color space into several sub-spaces where similar colors are associated together.

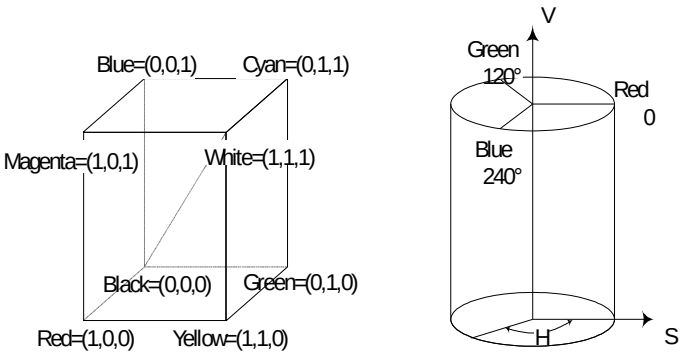

Fig. 4. (a) The RGB Color model (b) The HSV color model

The color values of the original pixels in an image are represented by the R, G, and B values, so that a transformation from the RGB to the HSV color model is necessary. It can be accomplished by the algorithm proposed.

IV. RESULTS AND DISCUSSION

A. Thesaliency Analysis

The main problem in determining a singular saliency map is adjusting for differences in magnitude and scale. However, simple normalization of each feature map would have the effect of enhancing low-level regions and reducing the importance of salient elements. The following procedure is implemented in order to give more importance to singular peaks than those that are repeated throughout the feature map:

- Normalizing all features to a range 0..1
- Determining the average value for each feature map
- Weighting by multiplying every element in the matrix by $(1-\mu)^2$, where μ is the average value in the map.

Once a saliency map has been obtained for each feature, two individual measures can be estimated:

Focus of Attention: The original saliency algorithm proposed by Itti, Koch has the objective of determining the focus of attention of an image and to simulate the serial process through which the image is scanned. The final saliency map is obtained through direct averaging of individual maps:

$$Focus\ of\ Attention = \frac{I_s + \frac{RG_S + BY_S}{2} + \frac{0°O_S + 45°O_S + 90°O_S + 135°O_S}{4}}{3}$$

- Complexity Map: A more useful calculation is obtained by performing an OR operation that registers the normalized maximum peak of each feature. This is achieved for each pixel at coordinates (x,y) by:

$$Complexity\ Map = \max\left[I_S(x,y), C_S(x,y), O_S(x,y)\right]$$

The final results for both approaches are included in the following figure for a set of input images.

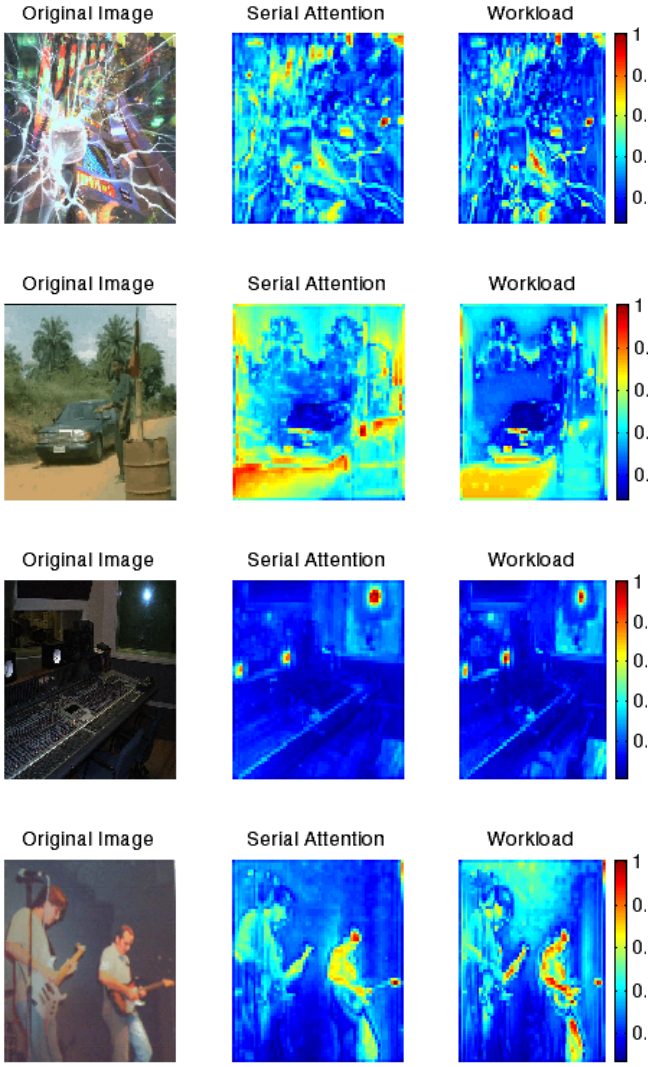

Fig. 5. Resulting Maps for thesaliency Analysis

B. Histogram Correlation

The histogram of an image is a measurement of the distribution of intensity/color values in a visual scene. It may be computed graphically by plotting pixel values along the horizontal axis and the number of occurrences of each value along the vertical axis (see figure.). The model calculates the correlation between adjacent histograms in the image.

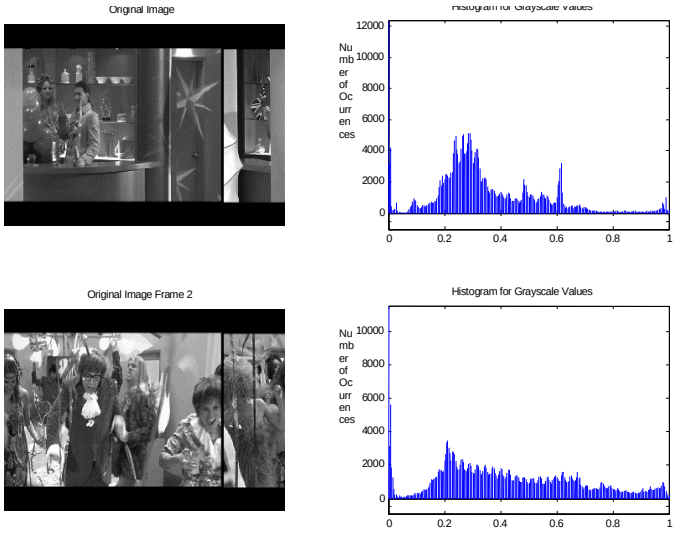

Fig. 6. Histogram representation of two frames in an image

C. Correlation Measurements

Correlation is a measure of the similarity of two signals for specific time lags. In the model, correlation is used to find similarities between frames. Seven correlation measures are used:

Direct Frame-to-Frame Correlation: In the case of two images, it indicates how much (or many) pixels from one frame differ from another. For similar frames, the correlation measure will approach 1 while for dissimilar frames the correlation measure will approach 0 (see figure).

Fig. 7. Correlation measures between sets of images

Summary of Visual Features

VIDEO FEATURE EXTRACTION VARIABLES						
	CLIP 1	**CLIP 2**	**CLIP 3**	**CLIP 4**	**CLIP 5**	**CLIP 6**
Correlation Frame to Frame						
Intensity	0.640	0.954	0.371	0.936	0.475	0.684
Histogram	0.986	0.986	0.742	0.999	0.843	0.838
Singular Values	0.999	1.000	0.995	0.998	0.996	0.999
DCT	0.848	0.981	0.808	0.970	0.770	0.862
FFT	1.001	1.001	1.009	1.000	1.063	1.006
Complexity	0.386	0.950	0.378	0.832	0.406	0.636
Focus of Attention	0.438	0.959	0.401	0.871	0.465	0.673
Standard Deviation						
Mean	0.233	0.221	0.097	0.211	0.202	0.060
Std	0.058	0.008	0.062	0.004	0.061	0.017
Euler Number						
Mean	-39.575	6.600	122.50	8.550	105.67	76.40
Standard	161.108	8.028	136.63	4.696	131.86	29.13
Centroid						
Std	23.139	5.742	55.679	7.968	96.889	41.14
Distance Mean	12.405	1.632	36.534	3.414	33.422	9.610
Complexity AVG						
Mean	0.230	0.224	0.215	0.213	0.185	0.238
Std	0.030	0.012	0.065	0.011	0.049	0.056
Complexity Median						
Mean	0.191	0.172	0.180	0.200	0.144	0.200
Std	0.034	0.011	0.067	0.012	0.054	0.057

D. FFT Correlation

The Fourier transform describes the frequency component of a given signal and for spatial 2-D signals is given by:

$$F(m,n)=\frac{1}{N}\sum_{k=1}^{N}\sum_{K=1}^{N} f(k,l)e^{-j\frac{2\pi}{N}(mk+ml)}$$

where m and n are the spectral coordinates, k and l are the spatial coordinates and N is the total number of samples for each dimension.

Fig 4 shows the model computes the correlation between adjacent FFT image transformations to account for frequency changes in the sequence of images.

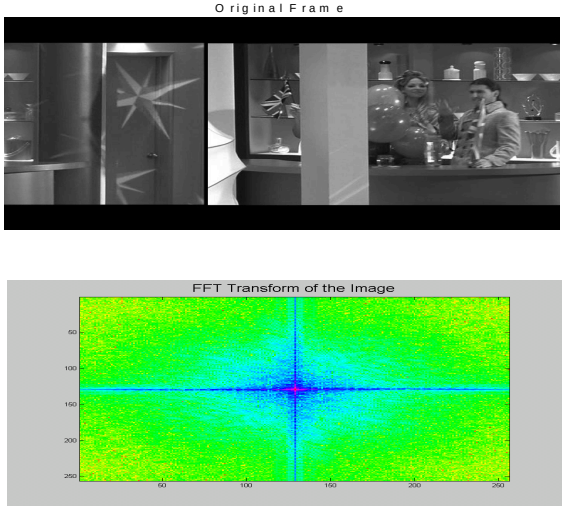

Fig 8. Relevant and Ir relevent images FFT Transformation of an image

V. CONCLUSION

This paper has proposed an automatic algorithm for harvesting the Web and gathering hundreds of images of a given query class. Thorough quantitative evaluation has shown that the proposed algorithm performs similarly to state-of-the-art systems such as while outperforming both the widely used Google Image Search and recent techniques that rely on manual intervention. Polysemy and diffuseness are problems that are difficult to handle. This paper improves our understanding of the polysemy problem in its different forms. An interesting future direction could build on top of this understanding as well as the ideas in [26] and leverage multimodal visual models to extract the different clusters of polysemous meanings, i.e., for tiger: Tiger Woods, the animal. It would also be interesting to divide diffuse categories described by the word airplane (airports, airplane interior, and airplane food) into smaller visually distinctive categories. Recent work addresses the polysemy problem directly and a combination with our work would be interesting.

Our algorithm does not rely on the high precision of top returned images, e.g., from Google Image Search. Such images play a crucial role, and future work could take advantage of this precision by exploiting them as a validation set or by using they directly instead of the text based ranker to bootstrap the visual training. There is a slight bias toward returning “simple” images, i.e., images where the objects constitute large parts of the image and are clearly recognizable. This is the case for object categories like car or wristwatch, where an abundance of such images occurs in the top text-ranked images. For other object classes, more difficult images are returned as well, e.g., elephant. The aim to return a more diverse set of images would require additional measures. Although some classification methods might require difficult images, gives an example of how a car model can be learned from these images. This automatically learned model is able to segment cars in unseen images.

REFERENCES

[1] J. Aslam and M. Montague, “Models for Metasearch,” Proc. ACM Conf. Research and Development in Information Retrieval, pp. 276-284, 2001.

[2] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan, “Matching Words and Pictures,” J. Machine Learning Research, vol. 3, pp. 1107-1135, Feb. 2003.

[3] T. Berg, “Animals on the Web Data Set,” http://www.tamaraberg. com/animalDataset /index.html, 2006.

[4] T. Berg, A. Berg, J. Edwards, M. Mair, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth, “Names and Faces in the News,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2004.

[5] T.L. Berg and D.A. Forsyth, “Animals on the Web,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2006.

[6] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet Allocation,” J. Machine Learning Research, vol. 3, pp. 993-1022, Jan. 2003.

[7] C.K. Chow and C.N. Liu, “Approximating Discrete Probability Distributions with Dependence Trees,” IEEE

Information Theory, vol. 14, no. 3, pp. 462-467, May 1968.

[8] B. Collins, J. Deng, K. Li, and L. Fei-Fei, "Towards Scalable Data Set Construction: An Active Learning Approach," Proc. 10th European Conf. Computer Vision, 2008.