TVGraz: Multi-Modal Learning of Object Categories by Combining Textual and Visual Features

Inayatullah Khan, Amir Saffari, and Horst Bischof

Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{*khan,saffari,bischof*}*@icg.tugraz.at*

Abstract
Internet offers a vast amount of multi-modal and heterogeneous information mainly in the form of textual and visual data. Most of the current web-based visual object classification methods only utilize one of these data streams. As we will show in this paper, combining these modalities in a proper way often provides better results not attainable by relying on only one of these data streams. However, up to our knowledge, there is no publicly available dataset for benchmarking algorithms which use textual and visual data simultaneously. Therefore, in this work, we present an annotated multi-modal dataset, named TVGraz, which currently contains 10 *visual object categories. The visual appearance of the objects in the dataset is challenging and offers a less biased benchmark. In order to facilitate the usage of this dataset in vision community, we additionally provide a preprocessed text data by using VIPS (VIsion-based Page Segmentation) method. We use a Multiple Kernel Learning (MKL) method to combine the textual and visual features in a proper way and show improved classification and ranking results with respect to the using only one of the data streams.*

1 Introduction

Today, Internet provides a vast amount of multi-modal information in form of textual and visual (images or videos) data. Extracting useful information and concepts from these data streams facilitates various vital tasks, such as web query and classification. In this paper, we focus on the task of visual object category learning from web resources. In particular, we try to present for the first time a public dataset that contains the whole web pages including both images and textual data. As we will show later in the paper, using both visual and textual data streams improves the ranking and classification performance of the visual object categorization task, which is necessary for web-based image search utilities.

Traditional web-based image search engines, such as Google, Flickr, and Yahoo use mainly the textual modality for retrieval of images. Typical features used with these search engines include the surrounding text around an image, web page title, keywords, description, and image file name [18, 7, 9]. The main idea behind these methods is that the rich textual information is meaningful and may describe the visual contents of the images [18]. However, the results based on textual elements may not always be accurate and may contain irrelevant images with respect to the queried category.

There exist another branch of approaches, which try to learn a visual concepts from images available on the web by using only the visual features. These methods usually use text-based image search engines to collect the data and then process them only based on the visual content of the images. For

example, Fergus et al. [6] developed a model to learn object categories from images retrieved from Google image search. Schroff et al. [17] reranked the images based on the surrounding text before using only visual features for classification of images.

Other approaches for learning object categories use both text and images simultaneously. Berg and Forsyth [2] studied the problem of learning object categories from the web and focus on the animal's categories. Four cues are combined to determine the final classification of each image: nearby words, color, shape, and texture. The surrounding text is used for pre-clustering of the images. Another similar concept based on graph-theoretical framework for web image clustering is presented in [8, 16]. Morsillo et al. [14] proposed a model based on a probabilistic graphical model which combines the visual and textual features simultaneously.

Although the combination of textual and visual data has been shown to improve the image ranking and retrieval quality, there is no publicly available multi-modal dataset in our knowledge. Such a dataset is essential for on-going research on visual category learning from web resources. Therefore, in this paper, we present a dataset (named TVGraz) of visual object categories by including visual and textual elements of a web page. In order to facilitate the use of textual data for the vision community, we present the textual part of this dataset both in raw form and also in pre-processed format. Additionally, we present benchmark results on this dataset. We represent both, textual and visual features, by the bag-of-words model and use the Multiple Kernel Learning approach [15] to find the best combination of these modalities for the classification task. Our experimental results show that combining both streams, textual and visual, out-performs methods using only one of these streams and justify that such a combination is beneficial for web-based visual object learning methods.

In Section 2 we introduce the methodology for collecting the dataset, while in Section 3 we describe how textual and visual features are extracted and used by the Multiple Kernel Learning approach to classify images based on these information streams. We show experimental results in Section 4 and conclude the paper in Section 5.

2 Dataset

We create a dataset, named TVGraz[1], consisting of 4030 images and associated web pages, for 10 categories as listed in Table 1. The objective of multi-modal dataset is to provide a common means for evaluation of object categorization research based on text and vision. The different categories are selected from Caltech-256 [10] dataset. For each category we tried to retrieve the top 1000 results from Google image search[2] using the category name as the query and select medium image size. For each result we captured the image, the image file name, the web page containing the image and the image URL. We filter those images which are not accessible directly from their respective original URLs (because either the links do not exist or the website is protected). We also filter the result to remove empty images, images with missing text data, painting images, and line sketched images. The images are provided in their real form as downloaded from the internet in order to provide an unbiased and challenging dataset.

The manual labeling of the images based on their visual contents is not an easy task. The text words and images can be ambiguous [2], e.g, "butterfly" could refer to "butterfly insects" or "butterfly

[1]TVGraz dataset is available at http://www.icg.tugraz.at/Members/kahn/tvgraz

[2]http://images.google.com

Nr.	Category	Positive Instances	Negative Instances
1	brain	209	107
2	butterfly	305	131
3	cactus	217	116
4	deer	324	140
5	dice	272	142
6	dolphin	272	127
7	elephant	223	111
8	frog	333	189
9	harp	230	250
10	pram	207	125

Table 1: Categories in TVGraz Dataset.

valve" as well as "'butterfly shaped fish". It is also difficult to break this polysemy-like phenomenon automatically [2]. To solve these ambiguities we label the images manually. For eight categories named human brain, cactus plant, deer animal, elephant animal, dice object, pram, dolphin fish, and musical instrument harp, we label each image as positive if there exist at least a single true instance of the object; otherwise labeled it as negative. The frog animal and the butterfly insect are difficult categories to label due to ambiguities. We select these categories explicitly as we think that this could be easily solved by combining both text and visual features. The frog images obtained from the search engines contains frog shape like computer mouse, frog puppets, real frog, frog cartoon characters, frog shape like toys and many more. We label frog image as positive, if it has at least a single instance of the real frog animal otherwise labeled it as negative. Similarly, for butterfly insect we label the image as positive, if it contains at least a single real instance of a butterfly; otherwise it is labeled as negative. For each category randomly selected samples of positives and negatives are shown in Figure 1.

In addition to the original raw data and in order to provide an easy start-up for researchers in the vision community, we also provide a pre-processed form of the textual part of the database. In this respect, it has been shown that the surrounding texts of an image on a web page usually has an important connection to the semantic contents of the image. However, it is hard to clearly define the exact relevant text close to the image on a web page because of the rich content of the web page. A web page may contain various texts in surrounding of an image such as navigation, advertisement and contact information, which are neither related to the image nor to the topic of the web page. However, there exist some methods, which try to provide an automated solution for this problem, such as Window-based approaches [4] or VIsion-based Page Segmentation (VIPS) [5].

A Window-based approach uses a fixed length window to extract the text before and behind an image by treating the HTML source as a text stream [4]. This method is fast but might not be accurate because of the web page's structure discussed above. It is also difficult to define a fixed length window for every web page. The VIPS [5] method extracts the text close to the image on a web page by analyzing the tree structure of the web page based on its visual presentation [5, 4]. Each node in VIPS tree corresponds to a block and the segments obtained are more semantically aggregated. Each node is assigned a value indicating the coherency of the contents in the block based on its visual perception. Thus, it is easy to extract the text close to the image by locating the block containing the image. An example of a web page segmented by using VIPS is shown in Figure 2.

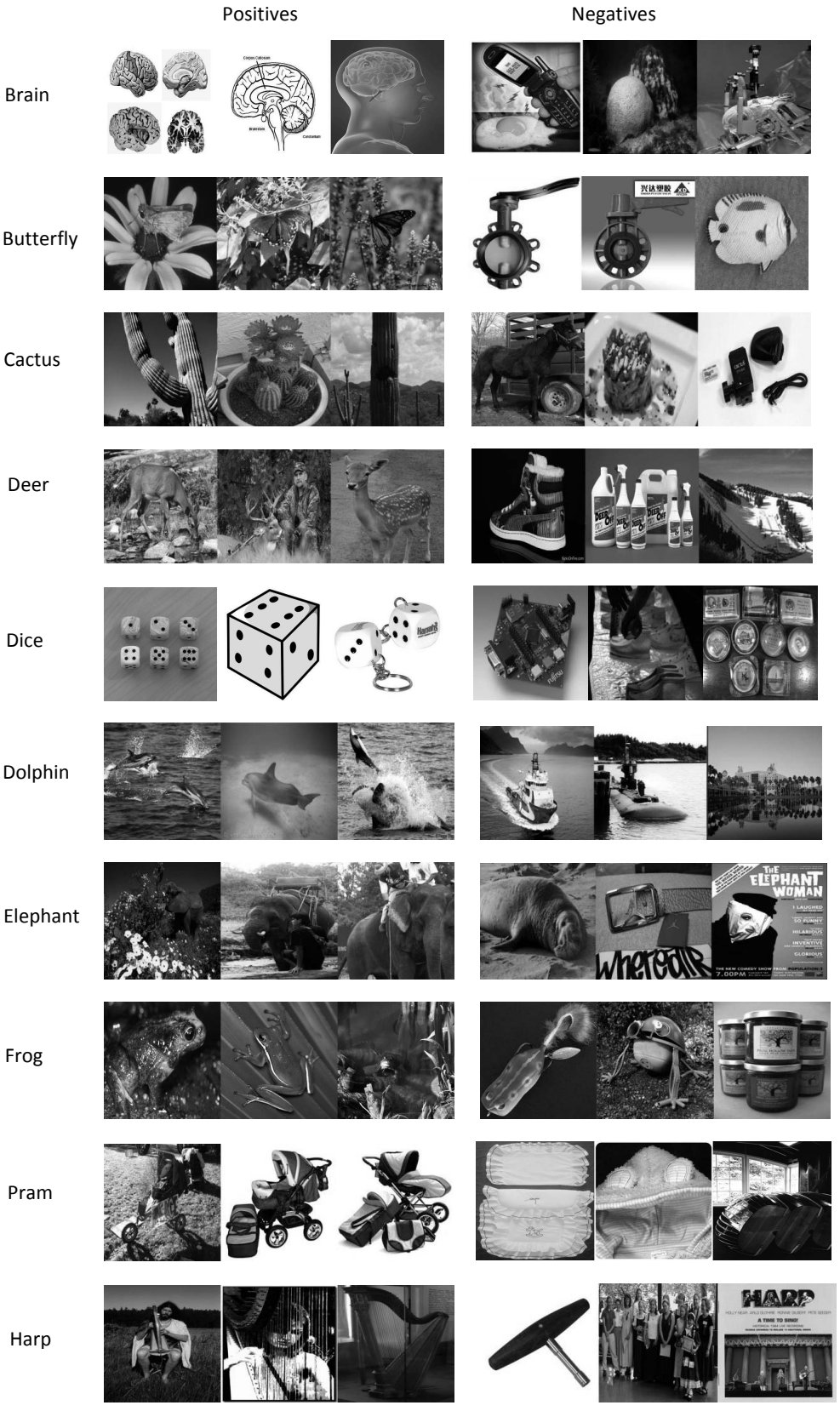

Figure 1: Sample images (positives and negatives) per category in TVGraz Dataset.

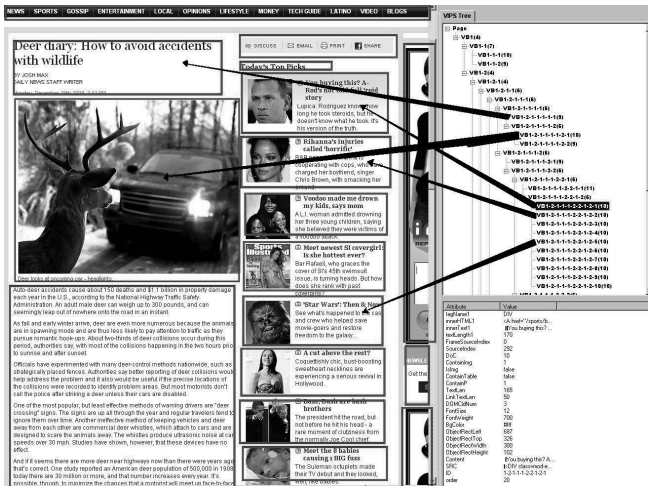

Figure 2: An example that shows VIPS based web page segmentation. The boxes show the visual blocks after segmentation. The text of the corresponding visual block may be used to describe the images contents

Therefore, we use VIPS for extraction of the text close to the image and additionally provide these as the pre-processed textual part of the dataset. For each image in the dataset, we use the image file name to locate the visual block containing the image after applying VIPS to the corresponding web page.

3 Learning from Visual and Textual Data

In order to combine textual and visual modalities, we need a proper representation of both features. We represent both features by the bag-of-words model. Multiple Kernel Learning method (MKL) is used to estimate weights of each feature in classification. In the following subsections, we provide the details for features representation and the Multiple Kernel Learning method.

3.1 Features Representation

For the textual data, we use the bag-of-words (BOW) model. It should be noted that the BOW model is a popular document representation scheme in text classification and information literature, where a document is represented as a bag of unordered words occurring in it. BOW is famous because of its simplicity and efficiency. In general, the text document is first parsed into words. Then a stop-list is used to remove insignificant words like 'with', 'wonder', 'the', and 'you'. Finally, words are represented by their stem or root by applying a stemming process, for example 'wait', 'waits', 'waited', and 'waiting' are represented by the root 'wait'. A unique identifier is then assigned to the remaining words and each text document is then represented by vector with components showing the counts of words it contains.

For each image, the textual features are built from the associated web page title, keywords, description, and all of the text close to the image. For each category all such text documents are parsed into words, then the number of words are reduced by applying the stop-list and stemming process. These words now form a dictionary and we represent the text data by the histogram of the word counts. As it has been mentioned before, we use VIPS [5] for web page segmentation; we use the TMG toolbox [21] for text processing.

For the visual feature extraction, we also use the standard visual bag-of-words model. Each image is converted to gray scale and resized to a 300 pixel width, keeping the same aspect ratio. We then apply a regular dense grid with 10 pixels spacing and extract SIFT descriptor [13]. Each grid point is represented by four SIFT descriptors, with circular support patches of radii 4, 8, 12, and 16. These multiple scale descriptors are used to provide a relative scale invariance. The dense descriptors are quantized into visual words using K-means clustering. The size of the codebook for each category is kept to 600, which is obtained from 50 randomly selected images positive images and 50 randomly selected negative images. Each image is then represented as a 600-dimensional histogram.

3.2 Multiple Kernel Learning

In the presence of multiple heterogeneous information sources, the Multiple Kernel Learning (MKL) approaches [15, 3, 12, 11, 1, 19, 20] provide a natural way to combine these data streams. The basic idea behind MKL is to create a weighted linear combination of the kernels from each information source, and adapt these weights in order to achieve the best performance.

Rakotomamonjy et al. [15] have shown that one can enhance the interpretability of the decision function by using multiple kernels instead of one. This provides flexibility in learning problems that involve multiple and heterogeneous data source, for example in our case involving text and vision. In such cases, a convenient approach is to consider the kernel $K(\mathbf{x}, \mathbf{x}')$ as a convex combination of the M basis kernels $K_m(\mathbf{x}, \mathbf{x}')$:

$$K(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{M} d_m K_m(\mathbf{x}, \mathbf{x}'), \tag{1}$$

where $d_m \geq 0$ are the corresponding weights of each source kernel, and we require that they sum to one: $\sum_{m=1}^{M} d_m = 1$. Each basis kernel $K_m(\mathbf{x}, \mathbf{x}')$ may either use the full set of variables describing $\mathbf{x}$ or subsets of variables stemming from different data source [12]. Therefore the decision function of an SVM with multiple kernels can be represented as

$$g(x) = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) - b = \sum_{i=1}^{N} \alpha_i y_i \sum_{m=1}^{M} d_m K_m(\mathbf{x}_i, \mathbf{x}') - b, \tag{2}$$

where $\mathbf{x}_i$ are the N training samples and $y_i \in \{-1, +1\}$ are the corresponding class labels. Learning both the coefficients α_i and the weights d_m in a single optimization problem is known as the main goal of the MKL methods [15]. After training, one can also analyze the kernel combination weights to estimate the relative importance of each information source for the classification task.

We use the approach introduced in [15], an efficient algorithm to solve MKL optimization problem by a primal formulation involving a weighted $l_2 - norm$ regularization. This method iteratively determines the combination of kernels by a gradient descent wrapping a standard SVM solver. In our case, the convex combination of the basis kernels becomes

$$K(\mathbf{x}, \mathbf{x}') = d_{txt} K_{txt}(\mathbf{x}, \mathbf{x}') + d_{vis} K_{vis}(\mathbf{x}, \mathbf{x}'), \tag{3}$$

where the subscripts vis and txt represent the visual and textual components, respectively. We use the χ^2 distance to form the visual kernel as

$$K_{vis}(\mathbf{x}, \mathbf{x}') = \exp(-\frac{D_\chi(\mathbf{x}, \mathbf{x}')}{\sigma}), \tag{4}$$

Category	Text	Vision	Text+Vision
brain	0.73	0.74	**0.81**
butterfly	0.72	0.71	**0.81**
cactus	0.78	0.76	**0.84**
deer	0.74	0.75	**0.80**
dice	0.67	0.61	**0.71**
dolphin	0.77	0.75	0.77
elephant	0.76	0.74	**0.79**
frog	0.64	0.73	**0.76**
harp	0.63	0.69	**0.73**
pram	0.73	0.77	**0.80**

Table 2: Average classification accuracy for all categories in TVGraz dataset. The cases where the combination of the textual and visual features result in a better performance are shown in bold cases.

where D_χ is the χ^2 distance and σ is the average χ^2 distances between all training samples. The χ^2 distance can be written as

$$D(\mathbf{x}, \mathbf{x}') = \frac{1}{2} \sum_{j=1}^{d} \frac{(x_j - x'_j)^2}{x_j + x'_j}, \tag{5}$$

where x_j is the j^{th} element of the vector $\mathbf{x}$.

The text kernel is represented as the linear kernel[3]:

$$K_{txt}(x, x') = x \cdot x'. \tag{6}$$

4 Results

In order to produce benchmark results on TVGraz dataset, we performed experiments based on three different trained classifiers. First a classifier is trained based on visual kernel, second with textual kernel, and finally with the combination of these kernels following the MKL approach. We trained a classifier per category by randomly selecting 50 positives and 50 negatives samples from the dataset. The remaining images were used for testing. In order to obtain a more robust estimate of the accuracy, we repeated these experiments 10 times and report the average performance.

Table 2 shows the average classification accuracy for each category. It is clear that in 9 out of 10 categories, we have a significant improvement by utilizing both information sources. The average precision-recall curves and the learned weight for each category are shown in the Figure 3.

Comparing the results from the textual and visual features, we can see that in some cases, the text part delivers a superior performance while in other cases vision performs better. However, when we combine these two data modalities, no matter which one is performing better, we can outperform their best individual result. This result encourages and confirms the idea that for a successful web-based visual object category classification system, using all available modalities would provide superior results.

[3]For text features we tried out different nonlinear kernels, but the performance of the linear kernel was comparable to the nonlinear kernels. Therefore for reason of efficiency we decided to use the simpler kernel

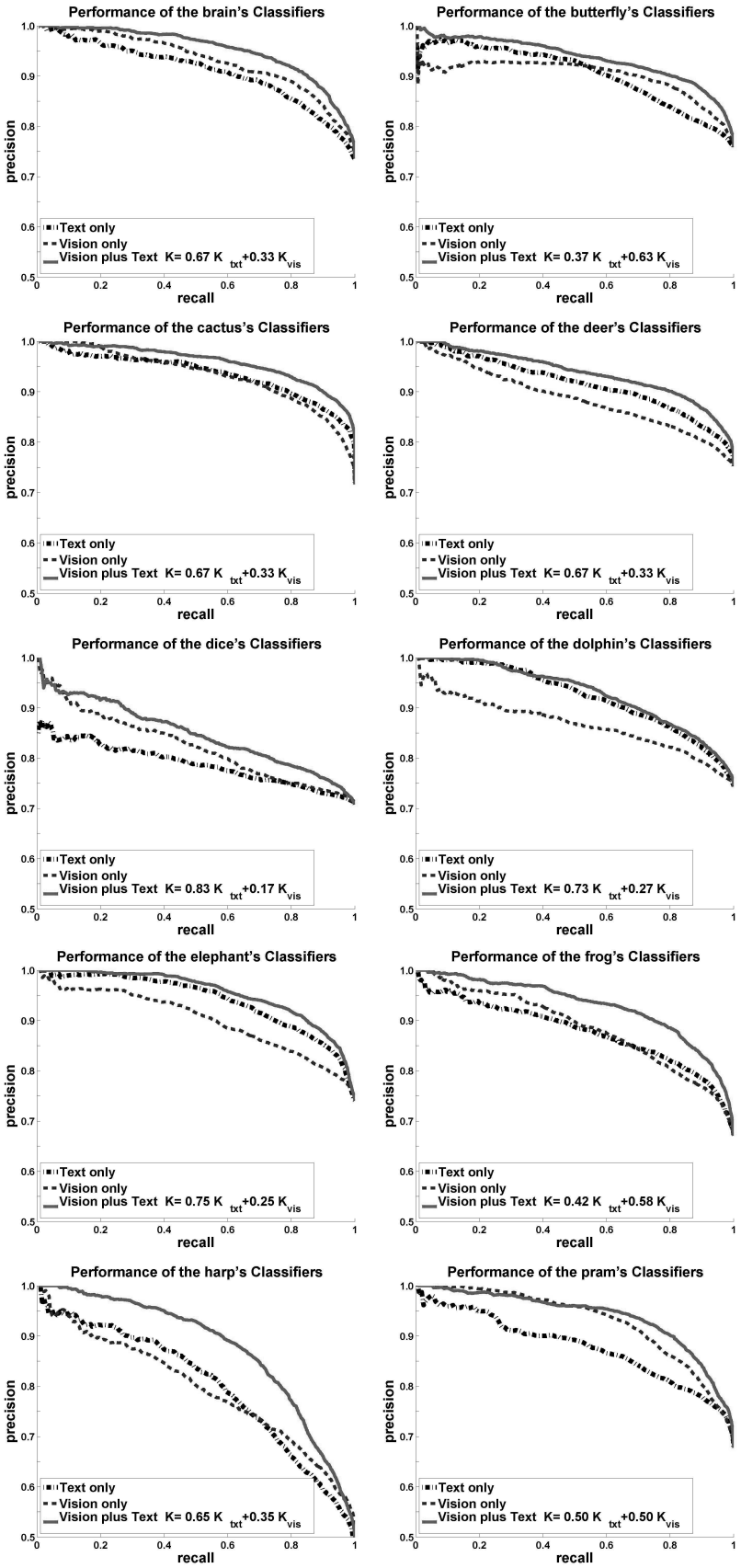

Figure 3: Average precision-recall curves and kernels weights for each category in TVGraz dataset. The black doted line shows the results based on textual features, the blue dotted line shows the results based visual features, and the red solid line shows the results based on combined textual plus visual features. For each category the weights for textual and visual based kernels are also shown.

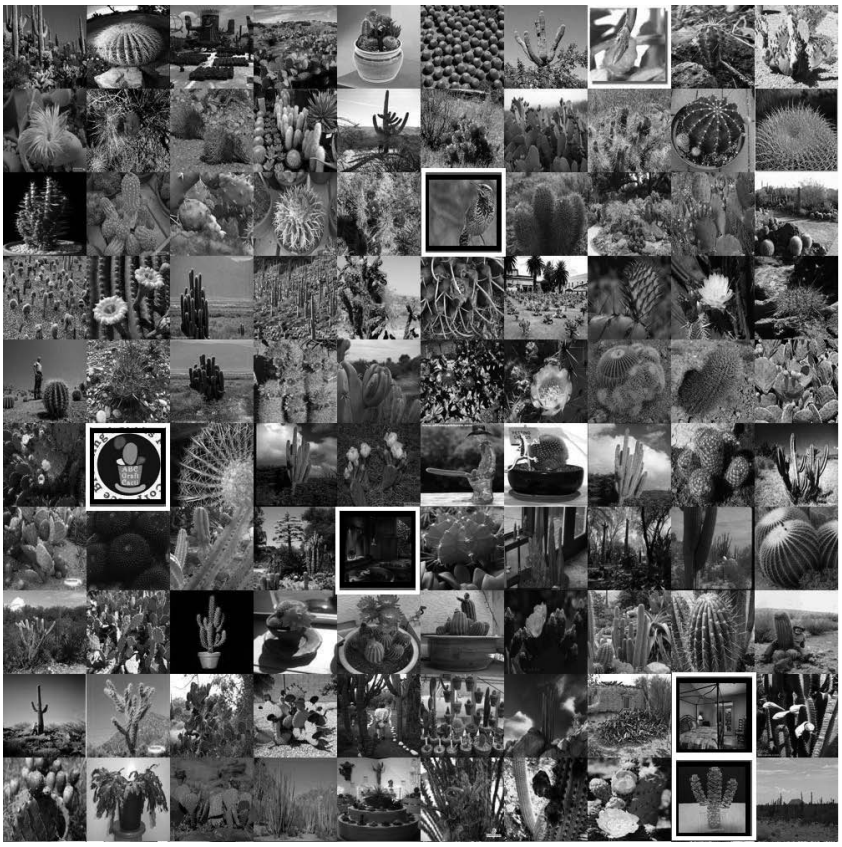

Figure 4: Top 100 images, ranking based on textual features for Cactus Class (white boxes show false positives).

In Figure 4, 5, and 6 we compare the top 100 images ranking and retrieval based on using text, vision, and their combination, respectively, for the Cactus class. We can see that the ranking of the test images based on the multi-modal learning removes many of the false positives and provides with a better search result.

5 Discussion and Conclusions

In this paper, we presented a public dataset (named TVGraz dataset) for benchmarking the web-based visual object category learning algorithms, which utilize both textual and visual information. We used a multiple kernel learning approach that provides a natural solution for learning from heterogeneous data sources. We showed that the combined strategy outperforms the methods using only one of the available data streams.

The idea of combining textual and visual features requires however the availability of both data sources. Therefore, for classifying images where the text data is missing or is very sparse, further steps and algorithms should be developed. For the task of web image search this is not a major issue; even filtering out those images without text information, there will be enough images left for the given task. Note, since current image search engines, like Google, use only the text information, the required textual data for these systems can already be provided.

Figure 5: Top 100 images ranking based on visual features for Cactus Class (white boxes show false positives).

6 Acknowledgment

This work has been supported by the Austrian Joint Research Project Cognitive Vision under projects S9103-N04 and S9104-N04 and by the Higher Education Commission of Pakistan under Overseas scholarships (Phase-II-Batch-I).

References

[1] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the International Conference on Machine Learning*, 2004.

[2] Tamara L. Berg and David A. Forsyth. Animals on the web. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 1463–1470, 2006.

[3] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the International Conference on Image and Video Retrieval*, pages 401–408, 2007.

[4] Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma, and Ji-Rong Wen. Hierarchical clustering of WWW image search results using visual, textual and link information. In *Proceedings of the International Conference on Multimedia*, pages 952–959, 2004.

Figure 6: Top 100 images ranking based on combined visual and textual features for Cactus Class with no false positives.

[5] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. VIPS: a VIsion-based Page Segmentation algorithm. Technical report, Microsoft Research, 2003.

[6] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from Google's image search. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1816–1823, 2005.

[7] R. Fergus, P. Perona, A. Zisserman, and Dept E. Science. A visual category filter for Google images. In *Proceedings of the European Conference on Computer Vision*, pages 242–256, 2004.

[8] Bin Gao, Tie-Yan Liu, Tao Qin, Xin Zheng, Qian-Sheng Cheng, and Wei-Ying Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *Proceedings of the International Conference on Multimedia*, pages 112–121, 2005.

[9] Zhiguo Gong, Leong, and Chan Cheang. Web image indexing by using associated texts. *Knowledge and Information Systems*, 10(2):243–264, 2006.

[10] G. Griffin, A. Holub, and P. Perona. The Caltech-256. Technical report, California Institute of Technology, 2007.

[11] Gert Lanckriet, Nello Cristianini, Peter Bartlett, and Laurent E. Ghaoui. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[12] Gert R. G. Lanckriet, Tijl De Bie, Nello Cristianini, Michael I. Jordan, and William S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[13] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.

[14] Nicholas Morsillo, Christopher J. Pal, and Randal Nelson. Mining the web for visual concepts. Technical Report 5272, University of Rochester Computer Science Department, 2008.

[15] Alain Rakotomamonjy, Francis R. Bach, Stèphane Canu, and Yves Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, November 2008.

[16] Manjeet Rege, Ming Dong, and Jing Hua. Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering. In *Proceedings of the International Conference on World Wide Web*, pages 317–326, 2008.

[17] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *Proceedings of the International Conference on Computer Vision*, pages 1–8, 2007.

[18] Heng T. Shen, Beng C. Ooi, and Kian L. Tan. Giving meanings to WWW images. In *Proceedings of the International Conference on Multimedia*, pages 39–47, 2000.

[19] Sören Sonnenburg, Bernhard S. Bernhard, P. Bennett, and Emilio Parrado-Hernández. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7, 2006.

[20] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[21] D. Zeimpekis and E. Gallopoulos. University of patras design of a MATLAB toolbox for term-document matrix generation, 2005.