WIKImage: CORRELATED IMAGE AND TEXT DATASETS

Doni Pracner [1]*, Nenad Tomašev* [2]*, Miloš Radovanović* [1]*, Dunja Mladenić* [2]*, Mirjana Ivanović* [1]

[1]Department of Mathematics and Informatics
Faculty of Sciences, University of Novi Sad
Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia

[2]Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
doni.pracner@dmi.rs, nenad.tomasev@ijs.si, radacha@dmi.rs
dunja.mladenic@ijs.si, mira@dmi.rs

ABSTRACT

This paper presents work towards the creation of free and redistributable datasets of correlated images and text. Collections of free images and related text were extracted from Wikipedia with our new tool *WIKImage*. An additional tool – *WIKImage browser* – was introduced to visualize the resulting dataset, and was expanded into a manual labeling tool. The paper presents a starting dataset of 1007 images labeled with any combination of 14 tags.

The images were processed into a number of scale invariant (SIFT) and color histogram features, and the captions were transformed into a bag-of-words (BOW) representation. Experiments were then performed with the aim of classifying data with respect to each of the labels on dataset variants with just the image information, just the textual data, and both, in order to estimate the difficulty of the dataset in the context of different feature spaces. Results indicate improvements in precision, recall and the F-measure when using the combined representation with *support vector machines* as well as the *k-nearest neighbor* classifier with the cosine similarity measure.

1 INTRODUCTION

A large body of research exists on mining image [3], and textual data [2], making these topics mature areas of research and practice. In addition, there exist many sources where one can find pictures with short captions associated to them, which can offer better understanding of the context underlying the images. In this paper we describe initial efforts to create correlated image and text datasets that will facilitate further research into these areas. One of the main guidelines in the creation of such datasets is to compose them out of free and redistributable images, thus enabling their free use for research purposes.

One of the options for collecting image data is crawling the Web, fetching images and surrounding text. This approach can be somewhat problematic since web pages tend to be chaotic, with a general lack of adequate context information. There are also numerous copyright issues: pictures are rarely labeled, and even when they are, the manner of labeling is not standardized.

Our approach to getting around the aforementioned issues is by crawling Wikipedia. Images are used in articles, thus captions for them can easily be extracted. Also, there exist categorizing schemes in Wikipedia. Most importantly, a lot of Wikipedia material is free to use and is labeled adequately, since the project makes efforts to protect itself from copyright infringements.

The rest of the paper is organized as follows. In the next section the new tool *WIKImage* is presented. Section 3 contains the description of an example dataset, as well as the feature representation. The results of classification experiments that were performed on the dataset are discussed in Section 4. Finally, in Section 5 a short summary, conclusions and future work options are given.

2 WIKImage

Our project for collecting image and text data from Wikipedia is named *WIKImage* – as a combination of the terms *WIKIpedia* and *Image*. In this section we will describe the general approach and technical issues regarding the process of data extraction from Wikipedia.

The wiki package Mediawiki, used to create Wikipedia, has a complex web-based API for accessing and editing information, which enabled us to retrieve the images and text. Entry point is of the form `http://www.example.com/w/api.php`. It can return data as XML, YAML, JSON, WDDX, serialized PHP, or human readable PHP arrays. All of these can be viewed on-line in a browser by appending "fm" to the format name.

A list of options is given when the API is called without parameters. A number of actions can be performed. In the context of this work the most important was the action *query*. It gives information about pages, their templates, categories, revisions, internal and external links, image usage, etc. Being that we are interested in obtaining free images, the retrieval was image-centric. There are several page categories that contain only free images, which provided a good starting point.

The process of obtaining images and text is split into two steps that can be executed simultaneously: retrieval and pro-

cessing, since the latter is much faster, and can be performed in several ways. Retrieval starts from a category, and attempts to fetch (at least) a predetermined number of images that are actually used on at least one page. They are all downloaded to the *images* folder. The API is asked for a list of pages that use the image, limited to results in the 'articles' name-space (that is, no user pages, talk pages, templates...). In the retrieval step the whole article is saved in the *pages* folder. Data about the usage of the image is saved separately in the *data* folder. All file names use the unique IDs of pages (both images and articles), and in all cases the existing data will not be downloaded multiple times.

Once everything is downloaded, adequate pieces of the text are extracted for every image. One option is for complete paragraphs that contain the link to the image to be extracted, and all of them stored in the *text* folder under the image ID name. Another option is to extract the captions for the images, since most of them use the standard wiki markup for these purposes. An attempt is also made to recognize if the picture is a part of a gallery construct. These extracted lines are stored in the *captions* folder. More refinements are expected to be made, depending on what will be needed in future work.

For a visual presentation of the obtained information the *WIKImage dataset browser* was created (Figure 1). It is a PHP web page that shows the images and appropriate text pieces page by page, and can be used to explore the datasets on-line.

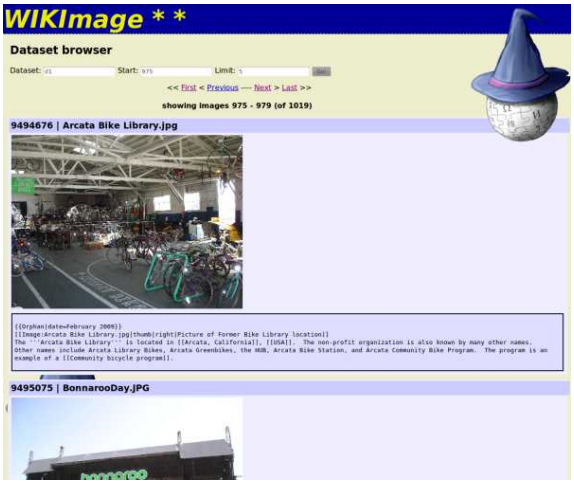

Figure 1: *WIKImage dataset browser.*

Our initial aim was to provide labels with all extracted data, to facilitate the task of automated classification. The original idea was to use Wikipedia's categories to automatically obtain class data about our images. But the problem with this is that the categories are chaotically applied to the images – some are not labeled at all, some are labeled with multiple categories. The categories themselves are very inconsistent in the level of detail. There are some hierarchies, but they are also problematic, since there are no central top-level categories. On the other hand, articles tend to be well categorized, but these categories are not necessarily relevant to the images used – for example in the article about Napoleon, there are images of him, but also of his tomb, the building where he lived on Elba, of the island itself, maps, battles, etc., all of which would not be appropriate to label with "people" or similar.

In the end, the decision was made to manually label the images into several binary categories, being that many images would be hard to classify into just one class. For this purpose the dataset browser was expanded into a dataset labeler to assist in the process.

3 FEATURE REPRESENTATION

Using the described tools, several datasets were obtained. We will present the dataset internally named "d1." For it's creation the English Wikipedia was used (API: `http://en.wikipedia.org/w/api.php`), with images from "Category: Creative Commons Attribution-ShareAlike 2.5 images." The dataset contains 1007 instances.

All instances were manually classified into binary labels shown in Table 1, along with the number of images in the dataset that were labeled with each of the tags. Many of the images had two or more tags applied to them. Some of these labels were introduced with the idea of being "secondary" – generally applied with another label, but with the potential for providing interesting results nevertheless. For instance, "sports" usually appears with "people" or "vehicles." Additionally, a few "special" labels were applied – for pictures with no captions, bad captions, and pictures that were found ambiguous.

Table 1: *Distribution of labels in the dataset.*

Label	Abbrv.	Instances
abstract and generated	abstract	6
animals	animals	89
art	art	53
buildings and constructions	buildings	375
documents and maps	documents	80
logos and flags	logos	43
machines, tools and tech	machines	23
misc nontech objects	misc	33
nature and scenic	nature	202
people	people	215
plants and fungi	plants	30
space	space	7
sports	sports	31
vehicles	vehicles	58

The image IDs, names, captions and label assignments (14 binary features) were exported to a comma-separated values (CSV) file for further processing.

Image analysis requires, first and foremost, an appropriate image representation which captures the information relevant for the task at hand. Many different feature types have been proposed, SIFT [5], GLOH [6] and LESH [9] being among the most frequently used. The basic idea is simple: detect some *keypoints* in images that are of interest for building the representation, and proceed by describing the local surrounding regions in a robust and detailed way. SIFT features in particular exhibit robustness to scaling, rotation and translation and are

hence very popular and used in many applications. This is why we opted for a SIFT-based representation of the acquired data.

More specifically, we used the *bag of features* representation, where groups of similar local features are treated as *visual words*. A *codebook* representing prototypical SIFT features was obtained by K-means clustering [4] on a large sample of features taken from all the images. The quantized representation is then obtained by mapping each individual feature to its closest prototype [11]. In our experiments (Section 4) we have used a 400-feature vocabulary. Since SIFT features are traditionally extracted from gray-scale images, we also appended a 16-bin color histogram to each SIFT representation [11].

As for the textual part of the representation, WEKA [10] software was used to transform the captions into a bag-of-words (BOW) representation. The *StringToWordVector* filter was used, applied with the alphabetic tokenizer which produces only "pure" words, resulting in a total of 3182 text attributes. No stemming was applied at this step, although it may be used in future experiments.

4 EXPERIMENTS

To demonstrate the difficulty of the constructed dataset "d1," and explore the influences of a combined representation, three datasets were created for each label – one with just the BOW representation (*text*), one with just the SIFT and histogram data (*image*), and one that contains both (*all*). All these representations are available for download at the WIKImage project home page [8].

The three created variants of data were tested in WEKA with the *k-nearest neighbor (kNN)* classifier (IBk implementation [1] with *Euclidean*, *Manhattan* and *cosine* distance/similarity measures) and a *support vector machine (SVM)* classifier (SMO implementation [7] with the default settings: $C = 1$ and the linear kernel), on binary classification problems corresponding to labels shown in Table 1.

The differences in percentage of accuracy between the image/text/all datasets were not great on average, with all of them scoring around 90%. There was a slight improvement when using all the features with SVM, as well as the cosine measure for kNN, but also a similar loss of accuracy with other distance/similarity measures.

What is more relevant for the observed imbalanced classification problems are the *precision*, *recall* and *F-measure* values, and here the differences were more notable, especially with SVM and kNN-cosine ($k = 1$) tests. Figure 2 overviews these results, showing averages of performance measures over all classification problems that were evaluated in 10 runs of 10-fold cross-validation. When using other distance metrics for kNN, the results for the text-only sets were much weaker than with the other sets. Likewise, usage of higher k values did not produce a notable difference in the results.

A more in-depth analysis of the results indicates that for most of the individual tags classifiers tend to be more successful with the combined datasets. Tables 2 and 3 show F-measure values (as the balance between precision and recall) for all the tags and types of datasets.

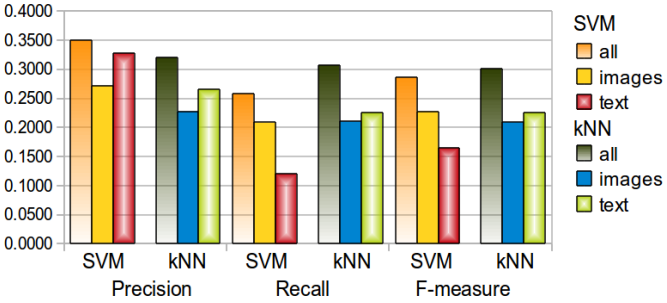

Figure 2: *Averaged results for SVM and 1NN (cosine).*

Table 2: *SVM – F-measure results.*

Dataset	All	Images	Text
abstract	0.0400±0.20	0.0000±0.00	**0.1000**±0.30
animals	0.2225±0.15	0.0825±0.10 •	**0.2659**±0.18
art	0.0961±0.14	0.0121±0.05	**0.1056**±0.15
buildings	**0.6807**±0.06	0.6262±0.06 •	0.4693±0.09 •
documents	**0.6334**±0.13	0.6161±0.14	0.2190±0.17 •
logos	**0.5844**±0.24	0.5807±0.26	0.0087±0.04 •
machines	0.0000±0.00	0.0000±0.00	0.0000±0.00
misc	**0.0999**±0.18	0.0925±0.16	0.0000±0.00
nature	**0.6603**±0.08	0.5976±0.09 •	0.3477±0.10 •
people	**0.5202**±0.08	0.4336±0.09 •	0.2394±0.10 •
plants	0.1253±0.22	0.0353±0.12	**0.1457**±0.23
space	**0.2000**±0.40	0.0900±0.28	**0.2000**±0.40
sports	0.0617±0.15	0.0000±0.00	**0.1272**±0.22
vehicles	0.0740±0.12	0.0217±0.07	**0.0752**±0.14
Average	**0.3014**	0.2091	0.2252

• statistically significant degradation compared to "all"

Table 3: *1NN (cosine) – F-measure results.*

Dataset	All	Images	Text
abstract	**0.2900**±0.45	0.0067±0.07	0.1800±0.39
animals	**0.4160**±0.14	0.1272±0.12 •	0.3346±0.20
art	**0.3216**±0.18	0.1222±0.14 •	0.2907±0.21
buildings	**0.5908**±0.05	0.5338±0.06 •	0.4990±0.19
documents	0.3891±0.15	**0.5151**±0.15	0.2444±0.17 •
logos	0.3118±0.18	**0.3359**±0.21	0.0241±0.07 •
machines	**0.0413**±0.14	0.0095±0.05	0.0375±0.13
misc	**0.0352**±0.11	0.0327±0.10	0.0029±0.03
nature	**0.4529**±0.09	0.4424±0.08	0.3748±0.16
people	**0.3856**±0.08	0.3666±0.08	0.3011±0.13 •
plants	**0.2491**±0.25	0.0020±0.02 •	0.2139±0.25
space	**0.3650**±0.46	0.1100±0.30	0.2900±0.45
sports	0.1457±0.17	0.1361±0.17	**0.1493**±0.19
vehicles	**0.2250**±0.17	0.1873±0.13	0.2110±0.19
Average	**0.2856**	0.2277	0.1646

• statistically significant degradation compared to "all"

Results shown in bold are the best for the corresponding row in the table, and the bullet signs next to the results denote statistically significant degradation compared to the combined representation, according to the corrected resampled t-test [10] at 0.05 significance level.

5 CONCLUSIONS

This paper presented work towards the creation of free and redistributable datasets of correlated images and text. Collections of free images and related text were gathered from Wikipedia by the *WIKImage* tool. An additional tool – *WIKImage browser* – was created to visualize the resulting dataset, and was expanded into a manual labeling tool. The paper presented an initial dataset of 1007 images labeled with combinations of 14 different tags.

The images were processed into a number of SIFT and histogram features, and the captions were transformed into a BOW representation. Binary classification tests were then performed on with respect to each of the labels, on dataset variants with just the image information, just the textual data, and the combination of both.

Initial experiments showed slight improvements in raw accuracy when using the combined sets for SVM and kNN with the cosine similarity measure. More importantly, there were notable improvements in *precision*, *recall* and *F-measure* values for these experiments, where on average the combined representation was the best and showed statistically significant improvements for several of the tags.

The datasets used in the presented experiments can be obtained from the WIKImage project home page [8].

Being that what this paper described are only the initial steps in the creation of the datasets, there is considerable space for future improvements.

The system for detecting captions and relevant text could be enhanced further. There are many examples of pictures used in info-boxes and similar templates, which (at this point) are ignored. Some of the more popular ones could be recognized automatically, producing more reliable captions.

Effort could be invested in searching for schemes for automatic assignments of classes to the images, based on the categories in Wikipedia, providing that suitable hierarchy schemes can be found. Similarly, the aforementioned Wikipedia info-boxes could be used to suggest classes for at least a part of the dataset, to reduce the manual effort in labeling. For instance, there are info-boxes for musicians, actors, etc., which could tag the pictures as "people," or info-boxes for states which could tag pictures as "logos and flags."

An interesting option for the future is to use the tool on other Mediawiki projects. The API we connect to can be changed in a single line – it can be used on any other Mediawiki page (for instance Serbian or Slovenian Wikipedia). This could make datasets of images with captions in different languages, producing various options for multi-language research.

The most important future steps concern dataset expansion. As of this writing, new and bigger sets are already being labeled. The labeling tool could be advanced to a cooperative project like Wikipedia, with tags submitted by different users.

Finally, more experiments with different algorithms for classification would be desirable, and also a more detailed comparison of the influence of various parameters. The BOW representation can also be tweaked in the future, for instance by using stemming rules.

Acknowledgments

This work was supported by bilateral project between Slovenia and Serbia "Correlating Images and Words: Enhancing Image Analysis Through Machine Learning and Semantic Technologies," the Slovenian Research Agency, the Serbian Ministry of Education and Science through project no. OI174023, "Intelligent Techniques and Their Integration into Wide-Spectrum Decision Support," and the ICT Programme of the EC under PASCAL2 (ICT-NoE-216886) and PlanetData (ICT-NoE-257641).

References

[1] D. W. Aha, D. Kibler, and M. K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[2] R. Feldman and J. Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007.

[3] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice Hall, 3rd edition, 2007.

[4] S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[5] D. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1150–1157, 1999.

[6] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[7] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, pages 185–208. MIT Press, 1999.

[8] D. Pracner. The WIKImage project home page. `http://perun.dmi.rs/pracner/wikimage/`, 2011.

[9] M. S. Sarfraz and O. Hellwich. Head pose estimation in face recognition across pose scenarios. In *Computer Vision Theory and Applications*, pages 235–242, 2008.

[10] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 3rd edition, 2010.

[11] Z. Zhang and R. Zhang. *Multimedia Data Mining: A Systematic Introduction to Concepts and Theory*. Chapman and Hall / CRC Press, 2008.