Object Recognition in WIKImage Data Based on Local Invariant Image Features

Nenad Tomašev *, Doni Pracner †, Raluca Brehar ‡,
Miloš Radovanović †, Dunja Mladenić *, Mirjana Ivanović † and Sergiu Nedevschi ‡

* Artificial Intelligence Laboratory
Jožef Stefan Institute, Ljubljana, Slovenia
Email: nenad.tomasev@ijs.si, dunja.mladenic@ijs.si
† Department of Mathematics and Informatics
Faculty of Sciences, University of Novi Sad, Serbia
Email: doni.pracner@dmi.uns.ac.rs, radacha@dmi.uns.ac.rs, mira@dmi.uns.ac.rs
‡ Computer Science Department
Technical University of Cluj-Napoca, Romania
Email: raluca.brehar@cs.utcluj.ro, Sergiu.Nedevschi@cs.utcluj.ro

Abstract—Object recognition is an essential task in content-based image retrieval and classification. This paper deals with object recognition in WIKImage data, a collection of publicly available annotated Wikipedia images. WIKImage comprises a set of 14 binary classification problems with significant class imbalance. Our approach is based on using the local invariant image features and we have compared 3 standard and widely used feature types: SIFT, SURF and ORB. We have examined how the choice of representation affects the $k$-nearest neighbor data topology and have shown that some feature types might be more appropriate than others for this particular problem. In order to assess the difficulty of the data, we have evaluated 7 different $k$-nearest neighbor classification methods and shown that the recently proposed hubness-aware classifiers might be used to either increase the accuracy of prediction, or the macro-averaged F-score. However, our results indicate that further improvements are possible and that including the textual feature information might prove beneficial for system performance.

Keywords: images, object recognition, local invariant features, WIKImage, Wikipedia, hubness, classification.

I. INTRODUCTION

Images are an important information source in many practical applications. Yet, this information is sometimes quite difficult to extract.

Standard image processing pipelines often generate vocabularies of visual words as typical local image features. Regardless of the resemblance, these representations are still more difficult to handle than actual text. In textual data, we can easily understand the semantics of individual words, while the visual words are not as easy to interpret. In annotated image data, this deficiency can be overcome by correlating the textual descriptions with the generated image features. WIKImage data [1] (http://perun.pmf.uns.ac.rs/pracner.wikimage/) was collected with that in mind. It is a set of publicly available images from Wikipedia (http://en.wikipedia.org/wiki/Main_Page) along with their captions and paragraph text. The project aims at eventually supporting cross-representational search via text-to-image and image-to-text queries.

Finding the best feature representation to represent the data is a difficult task. A recent study has shown that some image features are prone to hubness [2], an aspect of the dimensionality curse that skews the local relevance structure and poses severe difficulties in data analysis. We will discuss the emergence of hubs and its consequences in more detail in Section II-B.

A. Contributions

In this paper, we have compared three different types of invariant local image features on WIKImage data: SIFT, SURF and ORB. We have shown that different features affect the $k$-nearest neighbor structure in different ways, in accordance with previous observations outlined in [2]. The analysis reveals several interesting patterns regarding the semantic inconsistencies, neighbor occurrence profile purity and the overall skewness of the network of influence.

The experiments involving TF-IDF have shown no clear benefits on WIKImage data, which demonstrates that this popular weighting scheme is not universally beneficial in object recognition tasks.

We have performed a thorough evaluation of 7 different $k$NN classifiers on all feature types over a set of 14 binary classification problems defined by a presence or absence of certain types of objects in the images. The results imply that some of the recently proposed hubness-aware $k$NN classification models might be beneficial for improving system performance.

Additionally, we reveal some surprising consequences of class imbalance in WIKImage data where the minority classes play a highly negative role and their hubs have a negative impact on the classification process.

II. RELATED WORK

A. Similar Datasets

The main motivation behind the WIKImage project was the apparent absence of large datasets providing aligned textual and image data along with class affiliation labels.
Some datasets are created specifically for the purpose of testing a particular system, which does not support general-purpose experimentation. Most of the time they are not even published after the initial research [3]. Some research was done on using the existing aligned textual and image datasets and automatically labeling the instances by some learning approach [4].

A freely available dataset containing Wikipedia Pictures of the Day exists for studying topic learning models [5].

ImageCLEF Wikipedia Image Retrieval Datasets [6] represent a project most similar to WIKImage. The datasets are composed of images that are collected from Wikipedia and the text is extracted from the containing pages. They even offer images with captions in different languages where available.

However, the classes in ImageCLEF are handled differently to how it was done in WIKImage. The ImageCLEF datasets were built to test the efficiency of algorithms in detecting predefined topics for images. These are usually relatively specific descriptions of a set of pictures, such as “colored Volkswagen beetles,” “playing cards” or “white ballet dress.” The topic selection is done automatically from the logs of a Web query engine and the descriptions and sample images are provided as guides. Some of the data is available from the project site.

The dataset created for the ImageCLEF 2012 Scalable Image Annotation Using General Web Data Task is similar to the previous one [7]. The main difference is that it does not provide the actual images, since it was created by crawling the Web and cannot guarantee that there would be no copyright breaches. Therefore, it only offers the extracted image features (several variants of SIFT). There are annotations associated with the images, but of the 250000, only 3000 were manually labeled, and the rest were automatically assigned labels based on the calculated image similarities.

### B. Hubness

Many real-world networks are known to exhibit scale-free properties [8], as the node degree distribution approaches a power law. However, it was only recently discovered that the k-nearest neighbor occurrence distribution asymptotically approaches a power law as the intrinsic dimensionality of the data increases [9].

In real-world data, the occurrence distribution is usually still far from being scale-free, but it is often significantly skewed. This results in most points being either hubs or orphans (anti-hubs). Hubs are the centers of influence and, in this context, very frequent nearest neighbors. Anti-hubs, on the other hand, are the points that almost never occur in query result-sets.

The skewed distribution of influence is not necessarily a bad thing, as such networks are also known to be more robust to random noise [10]. The problem in this particular case is that hubs often act as semantic singularities and propagate semantically incorrect information throughout the network of influence [11] [2].

Indeed, neighbor hubs were first observed in a music retrieval and recommendation system [12] as frequently retrieved songs that were usually semantically unrelated to the queries. This resulted in suboptimal system performance. Dealing with hub songs and the skewed distribution of influence in audio data is still an ongoing research topic [13] [14]. The role of image hubs in image retrieval and object recognition applications has not yet received similar attention, though the results of a recent study suggest that several widely used local image feature types might be highly prone to hubness [2]. An image data visualization and assessment tool for examining the emergence of image hubs has recently been proposed [15].

As a consequence of a rise in the general awareness of the underlying issue, various novel machine learning and data mining methods have recently been proposed for learning under the assumption of high data hubness. This includes several hubness-aware clustering algorithms [16] [17], classification methods [18] [19] [20] [21], instance selection techniques [22], metric learning approaches [23] [24] and anomaly detection strategies [25]. Hubness-aware instance weights have also been successfully used for learning the appropriate semantic feature representation for cross-lingual document retrieval [26].

We will see that some of these hubness-aware methods can be successfully applied to the problem of object recognition in WIKImage data, increasing system performance. We will address these possibilities in Section V and Section VI.

### III. WIKIMAGE DATA

#### A. Forming the Dataset

WIKImage project’s primary goal was to create a labeled aligned dataset of correlated images and text that would be free to use in experiments – that is not burdened with copyright problems. Wikipedia is a free source of both textual and image data covering a variety of important topics. The WIKImage project relies on Wikipedia’s policies of assigning proper licenses to images, as they are known to follow proper copyright procedures. Wikipedia images are used in the articles, so the forming of an aligned dataset was possible by extracting the textual information from the articles where the images appear. This included the embedding paragraph as well as the caption. The extraction was done via the available Mediawiki API.

Defining a proper set of object categories was an important part of the project. We have considered applying some automatic labeling schemes, but have decided to manually label the dataset instead. This gave us better control of the quality of the process, even though it was certainly time-consuming. A system was developed for manual WIKImage image labeling, as shown in Figure 1.

The 14 binary class affiliation features are shown in Table I. Most of them are very general.

Some of the labels are basic – such as animals, plants and fungi or people. Some were introduced with the idea of being “secondary” – usually applied with some other. These are for instance sports and art, both of which will often be applied with people.

The D3 dataset used in this paper is the second public aligned textual and image dataset published by the WIKImage project. It was created from the Wikipedia Category Creative
Fig. 1. WIKImage system for manual image labeling

TABLE I

DISTRIBUTION OF LABELS IN THE WIKIMAGE D3 DATASET

<table>
<thead>
<tr>
<th>Label</th>
<th>Abbrv.</th>
<th>Instances</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>abstract and generated</td>
<td>abstract</td>
<td>875</td>
<td>7.61%</td>
</tr>
<tr>
<td>animals</td>
<td>animals</td>
<td>340</td>
<td>2.96%</td>
</tr>
<tr>
<td>art</td>
<td>art</td>
<td>568</td>
<td>4.94%</td>
</tr>
<tr>
<td>buildings and constructions</td>
<td>buildings</td>
<td>3018</td>
<td>26.26%</td>
</tr>
<tr>
<td>documents and maps</td>
<td>documents</td>
<td>1929</td>
<td>16.79%</td>
</tr>
<tr>
<td>logos and flags</td>
<td>logos</td>
<td>826</td>
<td>7.19%</td>
</tr>
<tr>
<td>machines, tools and tech</td>
<td>machines</td>
<td>358</td>
<td>3.12%</td>
</tr>
<tr>
<td>misc nontech objects</td>
<td>misc</td>
<td>972</td>
<td>8.46%</td>
</tr>
<tr>
<td>nature and scenic</td>
<td>nature</td>
<td>1639</td>
<td>14.26%</td>
</tr>
<tr>
<td>people</td>
<td>people</td>
<td>2287</td>
<td>19.90%</td>
</tr>
<tr>
<td>plants and fungi</td>
<td>plants</td>
<td>809</td>
<td>7.04%</td>
</tr>
<tr>
<td>space</td>
<td>space</td>
<td>63</td>
<td>0.55%</td>
</tr>
<tr>
<td>sports</td>
<td>sports</td>
<td>252</td>
<td>2.19%</td>
</tr>
<tr>
<td>vehicles</td>
<td>vehicles</td>
<td>931</td>
<td>8.10%</td>
</tr>
</tbody>
</table>

Commons Attribution-ShareAlike 3.0 images. A total of 15941 images were collected in the first run. Of these, 11491 were manually labeled with help from volunteers. A single label was applied to 8188 images, while the rest have multiple labels. Of these, most (2610) have only two labels.

B. Feature Representations

We compared three different types of local invariant image features in our experiments. We have extracted SIFT [27], SURF [28] and ORB [29] features from the WIKImage D3 data. These features are often used in object recognition, image matching, motion tracking and segmentation.

1) Feature extraction: SIFT – Scale Invariant Feature Transform has been introduced by [27]. SIFT features in WIKImage were extracted using the opensift library [30]. The first step in SIFT feature extraction is a search over a set of scales and image locations using a difference-of-Gaussian function for finding candidate interest points. At each candidate location a detailed model is fit to determine the exact location and scale. Keypoints are selected based on the estimate of their stability. Each keypoint location is assigned one or more orientations based on local image gradient directions. The last step extracts descriptors for each keypoint: the local image gradients are measured at the selected scale in the region around each keypoint.

SURF – Speeded Up Robust Features descriptors have been used by [28]. SURF is several times faster than SIFT and more robust against different image transformations. In order to extract the keypoints, SURF uses an approximation to the determinant of the Hessian blob detector. Around each point of interest the sum of Haar wavelet response is computed. The computation speed is achieved by the use of integral images (both for keypoint and for descriptor extraction).

ORB – Oriented FAST and Rotated BRIEF – features introduced by [29] are even more computationally efficient than SIFT and SURF, and they have similar matching performance, are less affected by image noise, and are capable of being used for real-time performance. The ORB algorithm detects FAST [31] points in the image and uses the intensity centroid [32] to measure the corner orientation. In order to extract descriptors for each keypoint, ORB uses an efficient rotation of the BRIEF [33] operator.

2) Codebook generation: For generating the bag-of-visual-words description of WIKImage images we have used the following generic algorithm for each of the 3 examined feature types:

- In order to find the central concepts in the local feature space, a feature sample was first obtained from the images. The sample was created by randomly selecting $N$ images from the original dataset. In the experiments presented in this paper, $N$ was set to 5000. The keypoints were extracted for each image and all the keypoints were integrated into the sample.

- K-means clustering was performed on each sample, forming 400 clusters. The resulting clusters represent the low-level visual concepts that define the codebook. Centroids were used to describe and model the individual clusters.

- For each image and each local feature, the closest codebook vector was determined. The number of local features belonging to each codebook cluster on each image was incrementally counted. These frequency counts determined the histogram that was used as the final codebook representation for each given image.

IV. LOCAL FEATURES AND EMERGING HUBS

We compared the bag-of-visual-words representations obtained from each invariant local feature type (SIFT, SURF, ORB) and estimated their impact on the k-nearest neighbor structure of the data. We have tried normalizing the visual words by TF-IDF, as it is a commonly used feature normalization approach in object recognition [34] [35].

A comparison of hubness as neighbor occurrence distribution skewness is given in Figure 2 for $k = 10$. It seems that the SURF feature representation exhibits the highest hubness on WIKImage data, closely followed by ORB. The SIFT
representation, on the other hand, exhibits a significantly lower neighbor occurrence skewness, even though anything above $SN_k = 1$ is still considered as high-hubness data. This shows that the choice of feature type might greatly affect the possible emergence of hub images in the data.

Another aspect of neighbor occurrences that affects the modeling of hub behavior within the hubness-aware classification models \cite{18,19,20} is the purity of the neighbor occurrence profile. Bad hubs that occur consistently within the neighbor sets of a single “wrong” class are easier to handle than those that are highly unpredictable and occur with close to equal probability in neighbor sets of various different classes. This can be measured by the reverse neighbor set entropy, as shown in Figure 4. A lower entropy means that the occurrences might be somewhat easier to model.

The comparison between the reverse neighbor sets of different feature types on WIKImage data in Figure 5 reveals a clear pattern. In all individual classification problems, SIFT features exhibit the highest reverse neighbor set entropy, followed by SURF and then only ORB.

V. SIMILARITY-BASED OBJECT RECOGNITION

Nearest neighbor methods are frequently used in classification, even though it is known that they often face some difficulties when dealing with high-dimensional data \cite{30,37}. Additionally, a recent research advocates the use of nearest neighbor methods in object recognition \cite{38}, though based mostly on object-to-class similarities. We will focus on a more typical $k$NN case where similarities are calculated between pairs of images instead. The $k$NN-based methods are often used for class imbalanced data classification \cite{39,40}, due to a high specificity bias. This might prove to be beneficial for handling the WIKImage data.

In order to estimate the difficulty of the WIKImage dataset, we have evaluated 7 different $k$NN classifiers on all binary classification problems defined on WIKImage data, for each compared feature representation. The basic $k$NN \cite{41} is a simple method that we have used as a baseline for other comparisons. The adaptive $k$-nearest neighbor \cite{42} (AKNN) adapts the neighbor size for each query based on the observed neighbor set. Different class weights are employed by NWKNN \cite{43} for better handling of class imbalance. We have also included 4 recently proposed hubness-aware $k$NN classifiers: hw-$k$NN \cite{44}, h-FNN \cite{18}, HIKNN \cite{19} and NHBNN \cite{20}.

All experiments were performed in 10 runs of 10-fold cross-validation. The $L_1$ metric was used to measure the distance between images. The neighborhood size of $k = 10$ was used in all experiments. All experiments were run both on the non-weighted feature representations as well as those weighted by TF-IDF. Due to the class imbalance, algorithm performance was evaluated by the macro-averaged $F_1$-score ($F_1 \text{macro}$) \cite{45}. The $F_1$ score for a particular class is defined as $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$ and represents the harmonic mean of precision and recall. We also considered the overall accuracy, for comparison.

As previously discussed in Section IV, it seems that TF-IDF feature weighting does not contribute to classification performance on WIKImage. This is demonstrated in Figure 5 for $k$NN, though a similar lack of improvement can be seen in other classifiers, as well. It is not surprising
that alternative feature weighting and selection schemes are currently being explored for object recognition, though there seems to be no consensus on what constitutes the state-of-the-art [46] [47] [48].

The neighbor occurrence models learned by the hubness-aware $k$NN classifiers increase the generalization capabilities of $k$NN. However, reducing the specificity bias is usually detrimental for learning under class imbalance [49]. As the WIKImage data is highly imbalanced, we hypothesized that the improvements in classification accuracy might be followed by a reduction in $F_{1}^{M}$ and vice-versa. Earlier studies have argued that NHBNN [20] is the most suitable hubness-aware classifier for class imbalanced classification [24], as it improves the $F_{1}$-score.

Indeed, as the case of people detection in WIKImage images (WM-l0) shows (Figure 6), hw-$k$NN, h-FNN, HIKNN and AKNN can sometimes increase the overall recognition system accuracy. However, they achieve a lower $F_{1}^{M}$ performance than the baseline $k$NN, as can be seen in Figure 7. On the other hand, NHBNN and NWKNN achieve a lower accuracy than $k$NN, but a significantly higher $F_{1}^{M}$ score. This holds in all considered feature representations.

Whether the accuracy or the $F_{1}^{M}$ is to be prioritized in practice, depends on the context and the domain. If a system were to implement both types of classification approaches, those with high specificity and those with better generalization – a user would be able to switch between different recognition modes, based on the current context. In class-imbalanced recognition tasks, $F_{1}^{M}$ is usually preferred, unless there are some a priori weights for penalizing certain types of misclassification.

A comparison between the classifier performance on different feature representations (Figure 7) shows that the best overall results are achieved on SURF bag-of-visual-words representations and far worse on ORB BoW representations. This is consistent with our earlier observations in Section IV. The SURF BoW had induces the lower label mismatch ratio and had a lower reverse neighbor set entropy than the SIFT BoW. This explains the observed differences in classifier performance.

The best performance on WIKImage object recognition tasks in our experiments was therefore achieved by using the SURF bag-of-visual-words representation, without the feature weighting, in conjunction with the Naive Hubness Bayesian $k$-nearest neighbor classifier (NHBNN). The average performance leaves some room for improvements, which we intend to achieve by exploiting the textual information contained in captions and enclosing paragraphs.

VI. CLASS IMBALANCE IN WIKIMAGE DATA

Learning under class imbalance is one of the major challenges in supervised learning systems [49]. The majority class often overwhelms the minority class in borderline regions and induces misclassification of the minority points due to a significant difference in relative class densities in the sample. However, this intuitive simplification does not always hold in high-hubness data [2]. This is confirmed by our analysis of the WIKImage data.
In the case of people detection (WM-10), the majority class comprises 79.5% of images in the data and 20.5% of images belong to the minority class, where people exist in the images. In this particular case, this ratio is reflected rather well in the distribution of hubs and anti-hubs between the majority and the minority class, as shown in Figure 8 and Figure 9.

Here we have defined hubs specifically as those images that occur more frequently than the threshold which is set at the two standard deviations above the mean (the neighborhood size $k$). This definition was used in [9] [17]. The same number of the least frequently occurring points is taken as anti-hubs.

There seem to be some minor differences in the exact ratio, depending on the feature type – but the majority class has the majority of hubs and the majority of anti-hubs in this particular recognition task. Therefore, the average influence of an image is not class-dependent in this case, as the total influence of each class is roughly proportional to the number of images it contains. However, there are some surprising differences in the nature of influence the images exhibit.

Figure 10 shows the ratio between the numbers of major bad hubs in the minority and the majority class on WM-10. Those are the hubs that induce the largest number of label mismatches in $k$-neighbor sets. The situation is effectively reversed. We can see that most bad hubs are minority hubs and that images that belong to the minority class exhibit a much more detrimental average influence than those of the majority class.

The consequences of the minority bad hub dominance can be seen in the confusion matrix of $k$NN classification for this particular task. On average, taken over all 100 test folds on the SURF BoW, there are 7831.8 true negatives (TN), 940 true positives (TP), 1295 false positives (FP) and 846.2 false negatives (FN). Therefore, the $\frac{FP}{FN}$ ratio is only 1.53, while the ratio between class sizes is 4. The minority class accounts for a surprising amount of misclassification, because of the presence of bad hubs. Even though there are more bad minority hubs in SIFT BoW, the $\frac{FP}{FN}$ for SIFT is about 2.70, due to a lower overall hubness of the SIFT BoW representation on WIKImage.

We also considered one point type characterization scheme that was recently proposed for analyzing the $k$NN structure and problem difficulty under class imbalance [50]. The neighborhood $k = 5$ is observed and points are marked as safe if 4 or 5 of their neighbors belong to the same class, borderline if 2 or 3 neighbors belong to their class, rare if only one point belongs to their class and outliers if none of their neighbors has their label.
Figures 11 and 12 show the distribution of point types within the majority and the minority class, respectively. There are big differences between these distributions, which also seem to be affected by the choice of feature representation.

Surprisingly, using the SURF BoW somewhat decreases the number of safe (interior) points in the majority class, compared to SIFT and ORB. However, as can be seen in Figure 12, it compensates for this by significantly increasing the number of safe and borderline minority points and decreasing the number of rare and outlier minority points. This is beneficial for minority class classification.

These experiments suggest that there is a subtle interplay between hubness of the feature representations and the class distribution imbalance. These correlations need to be assessed and taken into account when designing robust object recognition systems. We hope that these results will provide a good starting point for further analysis.

VII. CONCLUSIONS AND FUTURE WORK

Object recognition from image data is a difficult task relying mostly on a good and robust feature representation. We compared several types of local invariant image features: SIFT, SURF and ORB on 14 binary object recognition tasks on WIKImage D3 data.

As images are intrinsically high-dimensional, we have examined the hubness of all considered feature representations, as a newly described aspect of the dimensionality curse related to the skewed distribution of influence in the \( k \)-nearest neighbor topology.

The lowest overall hubness was present in the SIFT bag-of-visual-words representation, while both SURF and ORB representations exhibited very high hubness. However, the SURF BoW was shown to induce the lowest label mismatch rate in \( k \)-neighbor sets, which later resulted in the best average classifier performance. The ORB BoW had the highest label mismatch rate and the lowest object recognition performance.

We have examined the influence of TF-IDF weighting and concluded that it offers no benefits on WIKImage data in any of the considered quantized local image feature representations.

We have evaluated seven different \( k \)NN classifiers: \( k \)NN, AKNN, NWKNN, hw-\( k \)NN, h-FNN, HIKNN and NHBNN. The Naive Hubness Bayesian \( k \)NN (NHBNN) achieved the highest \( F^M \) score in the experiments, closely followed by NWKNN. However, these two methods had a somewhat lower accuracy than the other methods, so HIKNN or AKNN are to be preferred in cases where the accuracy is prioritized.

Our analysis reveals that the distribution of bad hubness plays an important role in determining the misclassification rate and that it should be taken into account in system design.

In future work, we intend to learn from these observations and create a robust object recognition and image retrieval system built on top of the WIKImage data. We also intend to include the textual information in the feature representation and learn the correlations between textual words and visual words represented by typical local image features.

Additionally, we will continue to work on developing new robust hubness-aware machine learning and data mining approaches that could be used to improve systems based on analyzing image data.

ACKNOWLEDGMENTS

N. Tomasev and D. Mrdjenic thank Slovenian Research Agency for support through program Knowledge Technologies P2-0103. D. Pracner, M. Radovanovic and M. Ivanovic gratefully acknowledge the support of the Serbian Ministry of Education, Science and Technological Development through project no. OI174023.

REFERENCES


