Stopwords Detection in Bag-of-Visual-Words: The Case of Retrieving Maya Hieroglyphs

Edgar Roman-Rangel and Stephane Marchand-Maillet

CVMLab, University of Geneva, Switzerland

Abstract. We present a method for automatic detection of stopwords in visual vocabularies that is based upon the entropy of each visual word. We propose a specific formulation to compute the entropy as the core of this method, in which the probability density function of the visual words is marginalized over all visual classes, such that words with higher entropy can be considered to be irrelevant words, i.e., stopwords. We evaluate our method on a dataset of syllabic Maya hieroglyphs, which is of great interest for archaeologists, and that requires efficient techniques for indexing and retrieval. Our results show that our method produces shorter bag representations without hurting retrieval performance, and even improving it in some cases, which does not happen when using previous methods. Furthermore, our assumptions for the proposed computation of the entropy can be generalized to bag representations of different nature.

Keywords: Bag-of-words, stopwords, retrieval, archaeology, hieroglyphs.

1 Introduction

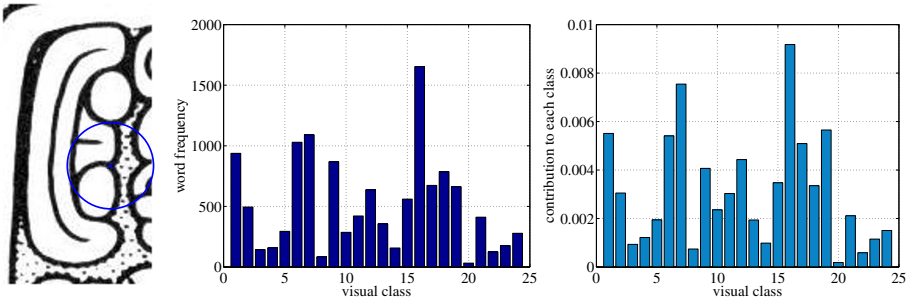
The bag-of-visual-words (BOW) paradigm [9] stands among the most efficient representations of images for the purpose of classification and retrieval [5] [12]. Based upon an analogy with the representation of text documents, this approach models a given visual document (i.e., image) as the frequency histogram of its *visual words*. In practice, the BOW model consists in a single vector per image which corresponds to a global representation constructed from a set of local descriptors, often robust to a variety of transformations, e.g., rotation, scale, affine, etc. As such, the BOW representation enables the capabilities for efficient indexing of large datasets and fast comparison of documents, as opposed to the expensive point-to-point matching paradigm.

One important step to generate discriminative BOW models of text documents consists in removing all stopwords from the vocabulary, i.e., words that are of low relevance for the BOW representation, e.g., articles or prepositions. This step is relatively trivial as text documents have finite vocabularies that can be easily discovered by identifying all unique words within the dataset. However, this is not the case for visual documents whose local descriptors are vectors in the continuous feature space R^N, and whose vocabularies often need to be estimated by the unsupervised quantization of the local descriptors [5]. As a consequence of

A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 159–168, 2013.

this, it is rather difficult to associate visual words to lexical components such as nouns or verbs, and consequently, equally difficult to decide which visual words should be considered stopwords and removed from the vocabulary.

Following the analogy with text representations, many past works have defined lists of visual stopwords as those words occurring with the highest frequencies [1] [9]. However, there is neither empirical nor statistical evidence that all the popular words in a corpus correspond to stopwords. Furthermore, recent works have shown that removing frequent visual words can harm the classification performance [11] [12]. Fig. 1 shows one example of a frequent visual word that is of high relevance and therefore must not be removed from the vocabulary.



(a) ©AJIMAYA. (b) Frequency per class. (c) Contribution per class.

Fig. 1. Example of a frequent visual word in a vocabulary of 1000 words estimated for a dataset of Maya hieroglyphs. 1a shows a visual example of the word, i.e., the blue circle indicates the local feature computed using the HOOSC descriptor [6]; 1b shows the frequency of this word across 24 visual classes; and 1c shows the contribution of the word to each class, proportional to the rest of the vocabulary. Although this word presents high occurrence count in the dataset, it is relevant as it distributes differently among the visual classes, thus contributing to the discriminative potential of the BOW representations.

In this work, we propose to use the probability distribution of each word as an indicator of the relevance of each such a word. More specifically, we consider that irrelevant words are roughly equally present in all visual classes, i.e., their probability distribution is uniform, thus resulting in high entropy [7], and therefore they must be considered stopwords. We also propose a specific formulation to compute the entropy as the core of this method, in which the probability density function of the visual words is marginalized over all visual classes.

We evaluate our methods on a dataset of syllabic Maya hieroglyphs that exists in binary format. This is a dataset of great importance for archaeologists devoted to the study of the ancient Maya civilization, as it serves the purpose of catalog to compare an identify new symbols as they are discovered. This dataset contains instances from 24 visual classes, while there are almost 1000 Maya symbols that have been identified thus far [6]. Yet it posits many challenges for automatic

retrieval. In practice, this dataset is a portion of a large corpus that is currently under construction, and that, in the long term, would benefit from automated methods that enable its efficient browsing and retrieval.

We compare our proposed approach with: (1) previous methods that simply consider stopwords those words having high or low frequency within a given dataset [1] [9]; and (2) a PCA-based method for dimensionality reduction [3] [2]. Our results show that different from these approaches, ours allows to remove up to 30% of visual words with almost no drop in the retrieval precision. Furthermore, it results in higher retrieval precision in some cases.

The rest of this document is organized as follows. Section 2 presents the related work in the detection of visual stopwords. Section 3 introduces our proposed approach. Section 4 explains the dataset and protocol followed to evaluate our proposed approach. Section 5 discusses our results. And section 6 presents our conclusions and potential future directions for this research.

2 Related Work

The most common approach to remove visual stopwords consists in simply excluding all the most frequent words from the vocabulary [1] [9]. This approach was devised based the analogy that text-stopwords have high frequencies over the whole corpus, e.g., articles and prepositions [9], and assuming that in the images they will point to local descriptors corresponding to background or simply to uninformative structures.

A more recent work [4] shows that the most frequent words are in fact not very informative, as the classification performance is not impacted by removing them. In addition, some new experiments show an increased performance after removing the most frequent words from the vocabulary [10]. However, some other recent explorations on this regard have shown that simply excluding the most frequent visual words may be harmful for the classification performance [12]. Furthermore there is not guarantee that all of the most frequent terms correspond to stopwords, and some claims have been raised in favor that frequent visual words are highly informative [11].

A related approach that pursuits dimensionality reduction of BOW representations is PCA [3]. This approach consists in projecting the BOW representations onto linear uncorrelated variables that are sorted according the amount of variance they are able to describe. These projected variables are called principal components, and a reduced BOW representation can be constructed by using only the N most principal components. However, this approach does not guarantee that the components excluded from the reduced representation correspond to stopwords. Also, the discussion regarding whether the loss in performance is acceptable or not remains open [2].

Upon an entropy-based approach, visual stopwords are automatically detected in a video corpus [13], where words correspond to actions and their distribution is conditioned to their occurence within each video. However, this approach suffers from two main drawbacks: (1) the list of actions (words) must be manually

defined, which might be unfeasible when there are no labels associated to the visual content; and (2) the entropy of each visual word is sensitive to small variations in its density across videos.

Similar approaches to ours are used to remove stopwords from text corpora [14] [15] and web documents [8], with the difference that the distribution of words is conditioned directly over the documents rather than on the classes, thus these approaches are sensitive to intra-class variations. However, to the best of our knowledge, there are not previous works that detect visual stopwords for still images based on their entropy.

3 Identifying Stopwords

This section explains our proposed method to compute entropy for visual words.

The entropy [7] can be used as an indicator of information content for a given random variable X, in the sense that the entropy of this variable will be larger as its probability density function (PDF) tends to the uniform distribution. This follows after the observation that little information is represented when all the possible values of a random variable are equally probable, i.e., no information is acquired by knowing the value of the random variable. Based on this observation, we define a visual stopword as a word with high entropy.

The Shannon entropy [7] of the discrete random variable X, with possible values $\{x_1, x_2, \ldots, x_n\}$ is defined as,

$$H(X) = -\sum_{i=1}^{n} p\left(x_i\right) \log p\left(x_i\right), \tag{1}$$

where $p\left(x_i\right)$ denotes the PDF of the random variable X, i.e., the probability of the variable taking the value $X = x_i$.

For the case of visual words, we marginalize their PDF over all possible visual classes in the dataset. This is, we can consider each visual word as a random variable W that can take values $\{w_1, w_2, \ldots, w_n\}$,

$$w_i = P\left(W|c_i\right), \quad i = 1, \ldots, n, \tag{2}$$

where c_i denotes each of the n visual classes in the dataset, and $P\left(W|c_i\right)$ indicates the relevance of W given the i-th class.

In practice, we can compute the PDF of the random variable W directly by,

$$p\left(w_i\right) = \frac{f^W\left(c_i\right)}{\sum_i f^W\left(c_i\right)}, \tag{3}$$

where $f^W\left(c_i\right)$ denotes the frequency of the word in class c_i, and the normalization induced by the denominator corresponds to the summation over all the visual classes within the dataset, but independently for each visual word.

Therefore, for each visual word in the dataset, the entropy of its associated random variable W can be computed as,

$$H(W) = -\sum_{i=1}^{n} p(w_i) \log p(w_i) . \tag{4}$$

After the entropy of each word is computed, we mark as stopwords all visual words with high entropy values, and remove them from the vocabulary.

4 Experiments

This section presents the dataset used to evaluate the proposed approach for detection of visual stopwords, and the experimental protocol we followed.

4.1 Dataset

The dataset consists of 24 visual classes of syllabic Maya hieroglyphs in binary format. Each class having 10 instances manually segmented from larger inscriptions that were collected by the AJIMAYA project [6], and that correspond to manual drawings traced upon photographies of stone monuments. This dataset also contains 25 synthetic variations that were randomly generated for each instance with the purpose of increasing the size of the dataset, thus accounting for 260 final instances per visual class.

The process to generate the synthetic data consisted in shifting the position of the bounding box containing the glyph of interest within the inscription that contains it, but without modifying its original size. During this process, the position was shifted at fixed intervals of $\{-16, -8, 0, 8, 16\}$ pixels from the original position, both in the horizontal and vertical axes. Therefore, the synthetic variant contain only sections of each original instances plus random visual structures that are present in the large inscription. Fig. 2 shows some examples the instances in the Maya dataset.

4.2 Experimental Protocol

We compare the performance of our proposed method against two approaches. Namely, (a) the traditional method that removes the most and less frequent visual words [1] [9], and (b) a PCA-based method [3] [2] for dimensionality reduction which assumes that the less principal components correspond to stopwords. Independently for each of these methods, we performed retrieval experiments to evaluate the proposed approach under the following protocol:

1. Compute local descriptors for all the images. We used the HOOSC descriptors [6] for the binary shapes.
2. Estimate a visual vocabularies of different sizes using a random subset of the descriptors and the k-means clustering algorithm.

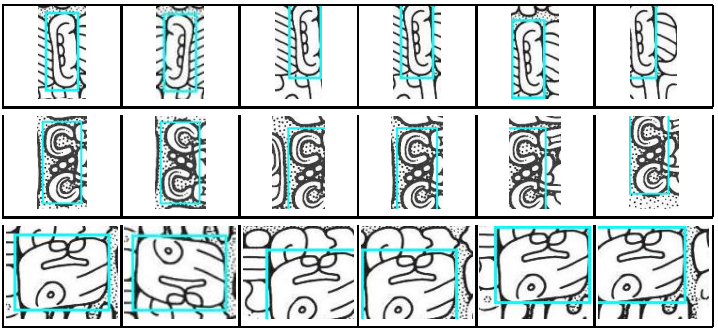


Fig. 2. Examples of Maya instances. The first element in each row is the initial segmented glyph, and the following elements are some of its synthetic variations. The glyph of interest is highlighted by a cyan rectangle in all cases. ©AJIMAYA

3. Estimate the BOW representation of each image, removing at each time, different portions of the vocabulary, i.e., the words that most likely are stopwords given the evaluated method.
4. Perform retrieval experiments under a leave-one-out full-cross validation, i.e., each of the instances in the dataset, one at the time, was considered as query, and the rest of the instances in the dataset were ranked according their visual similarity with respect to the current query. We use the Manhattan distance as ranking function, except for the PCA-based method that uses Euclidean distance since this metric is more suitable due to the Gaussian underlaying model and the orthonormality induced by the PCA projections.
5. Compute the mean Average Precision (mAP) of the retrieval experiment.

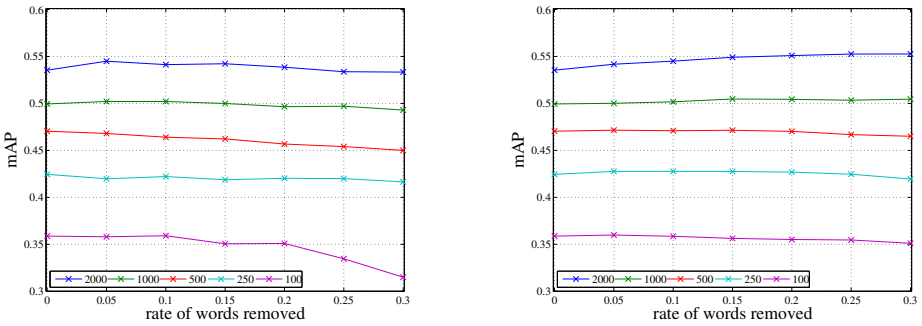
Note that the retrieval experiments were repeated several times, removing a different percentage of the visual vocabulary each time. All of our results are reported in terms of mAP.

5 Results

We start by evaluating the retrieval performance of the baseline approach that removes the most frequent visual words by varying the rate of words removed. Fig. 3a shows the mAP achieved by removing up to 30% of the most frequent words in dictionaries of different size. Note that the performance exhibits a slight degradation as more words get excluded, specially for the smaller dictionaries.

Similarly, Fig. 3b shows the mAP obtained after removing up to 30% of the visual words based on our method. Contrary to the frequency-based case, the performance only degrades for small dictionaries with the entropy-based approach, and provides a slight improvement for larger dictionaries. This result shows that it is possible to obtain compact dictionaries that not only hold good retrieval performance but also could improve it in some cases.

To validate the idea of considering words with low frequency to be stopwords as suggested by some previous works [1] [9] [10], we combined such an approach



(a) Frequency-based. (b) Entropy-based.

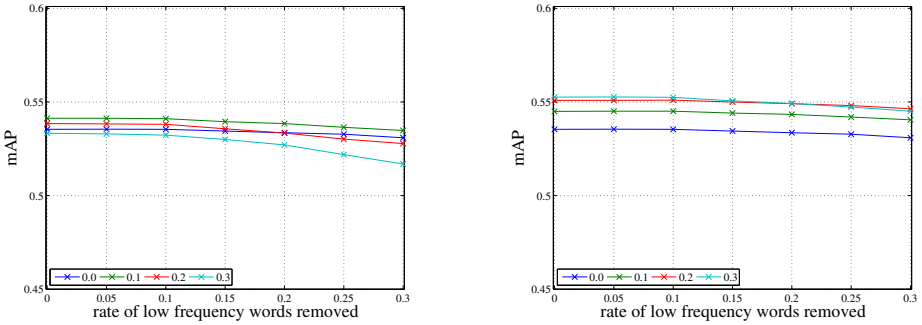
Fig. 3. *mAP* achieved by removing words from the BOW representations at different rates based on: 3a their high frequency; and 3b their high entropy. Each curve corresponds to a vocabulary of different size, i.e., 2000, 1000, 500, 250, and 100 visual words, respectively.

with both the baseline and our proposed approach. This is, we removed (a) words of high frequency and also words of low frequency first, and then (b) words of high entropy together with words of low frequency. Fig. 4a shows the degradation in the retrieval performance as more words of low frequency get removed from the BOW representations using the vocabulary of 2000 words, where each curve corresponds to a fix rate of words removed using the criteria of high frequencies. Note that all of these curves degrade as more words of low frequency are excluded. A very similar behavior is found when the approach of excluding the less frequent words is combined with the proposed approach based on high entropy, as shown in Fig. 4b. The curves obtained for the dictionaries of 1000, 500, 250, and 100 words all have similar behavior when removing words of low frequency (we do not show those curves here due to space limitations).

This consistent degradation in the performance suggests that words of low frequency do not correspond to stopwords. However, the performance is not drastically affected by removing up to 30% of these words. More specifically, by looking at the cyan curve in Fig. 4b[1], that corresponds to removing 30% of the words with higher entropy, we can see that it is also possible to remove up to 10% of the less frequent words without hurting the retrieval performance, thus 800 words in total. This means that there are roughly 1200 words that are of actual relevance for a vocabulary of 2000 visual words. Note that this does not imply that a direct estimation of a vocabulary of 1200 words would achieve the same performance, as this would change the structure of the vocabulary.

The final approach that we evaluated consists in a PCA-based dimensionality reduction. Fig. 5a shows the curves of the retrieval precision obtained after removing words that correspond to the less principal components. Note that

[1] Best viewed as pdf since the red curve overlaps the cyan curve.



(a) Frequency-based. (b) Entropy-based.

Fig. 4. *mAP* achieved by removing words of low frequency from the BOW representations at different rates, and using a fixed dictionary of 2000 visual words. Each curve corresponds to the approach that removes words based on: 4a their high frequency; and 4b their high entropy. Note that the two blue curves are the same, as they mean that words are excluded solely based on their low frequency.

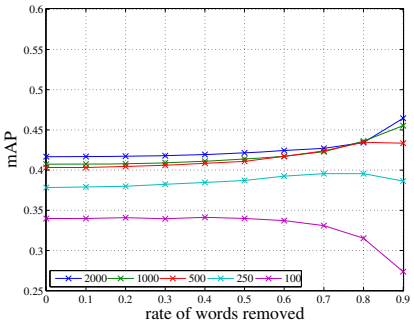
in this case, removing more words results in better retrieval precision for the largest dictionaries, e.g., the best performance is achieved by using only 10% of the words from the dictionaries of 1000 and 2000 words (blue and green curves). In Fig. 5b, the retrieval precision of PCA is further compared with an entropy-based method that also uses the Euclidean distance for ranking, red and blue curves respectively. We can see that these two approaches achieve similar results up to 70% of the words are removed, and that PCA is more suitable that the entropy to produce very short representations, although at the price of a small drop in performance when compared with the entropy-based method that uses the L1 distance (green curve, which corresponds to the green curve also in Fig. 3b). Note that this comparison corresponds to the vocabulary of 1000 words, and that the behavior remains similar for vocabularies of different sizes.

In practice, the proposed entropy-based approach proves to be a suitable option for detection of visual stopwords, as it ensures that only those words that are irrelevant for a given visual vocabulary get excluded from the final BOW representations, i.e., words having a uniform distribution over the visual classes.

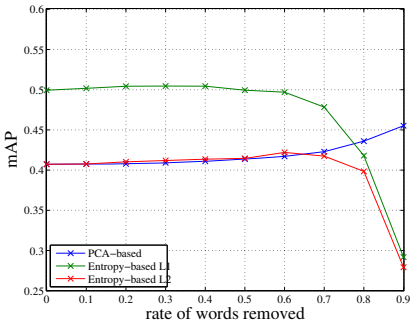
6 Conclusions

We presented a method to detect stopwords in visual vocabularies, and proposed a formulation to compute the entropy of visual words based upon their probability distribution, which is marginalized over all the visual classes but independent for each visual word.

We evaluated our proposed formulation on a dataset of syllabic Maya hieroglyphs, which consist in complex shapes. The dataset we used represents but



(a) PCA-based.



(b) PCA-based vs Entropy-based comparison for 1000 visual words.

Fig. 5. *mAP* achieved by removing words from the BOW representations at different rates based on the amount of variance they describe after applying PCA on the visual vocabulary

a portion of a potentially enormous corpus that its currently under construction, and which will require adequate representations for efficient indexing and automatic retrieval, thus the motivation for this research.

Our proposed formulation ensures that only those words that are irrelevant for a given visual vocabulary get excluded from the final BOW representations. This is those visual words having a uniform distribution over the visual classes, regardless of their distribution across a given document or their frequency within the whole dataset. In practice, our results demonstrate that the proposed formulation provides with shorter BOW representations of equal or higher discriminative power to retrieve complex shapes. Such shorter representations are equally capable of good retrieval results as the original representations constructed after simple quantization, and much more suitable than those found under a PCA-based dimensionality reduction approach. Furthermore, a slight improvement in the retrieval precision was possible in some cases.

It remain to be evaluated whether our findings will be true also for visual vocabularies of different nature, e.g., bag representations for intensity images, or based on other kind of local descriptors. However, we believe that our approach will be equally suitable for those cases, under the same assumption that the entropy can guide the detection of visual stopwords, as it previously did for the detection of text stopwords.

Acknowledgments. This works was supported by the Swiss NSF through the NCCR IM2.

References

1. Hsiao, J.-H., Chen, C.-S., Chen, M.-S.: A Novel Language-Model-Based Approach for Image Object Mining and ReRanking. In: Proceedings of the 8th IEEE International Conference on Data Mining (2008)
2. Jégou, H., Chum, O.: Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 774–787. Springer, Heidelberg (2012)
3. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2010)
4. Jiang, Y.-G., Yang, J., Ngo, C.-W., Hauptmann, A.G.: Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study. IEEE Transactions on Multimedia 12(1), 42–53 (2010)
5. Quelhas, P., Monay, F., Odobez, J.-M., Gatica-Perez, D., Tuytelaars, T.: A Thousand Words in a Scene. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(9), 1575–1589 (2007)
6. Roman-Rangel, E., Pallan, C., Odobez, J.-M., Gatica-Perez, D.: Analyzing Ancient Maya Glyph Collections with Contextual Shape Descriptor. International Journal of Computer Vision 94(1), 101–117 (2011)
7. Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal 27(3), 379–423 (1948)
8. Sinka, M.P., Corne, D.W.: Towards Modernised and Web-Specific Stoplists for Web Document Analysis. In: Proceedings of the IEEE/WIC International Conference on Web Intelligence (2003)
9. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of the 9th IEEE International Conference on Computer Vision (2003)
10. van Zwol, R., Garcia Pueyo, L.: Spatially-aware indexing for image object retrieval. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining (2012)
11. Yang, J., Hauptmann, A.: A text categorization approach to video scene classification using keypoint features. CMU Technical Report (2006)
12. Yang, J., Jiang, Y.-G., Hauptmann, A.G., Ngo, C.-W.: Evaluating Bag-of-Visual-Words Representations in Scene Classification. In: Proceedings of the International Workshop on Multimedia Information Retrieval (2007)
13. Zhao, Z.: Towards a Local-Global Visual Feature-Based Framework for Recognition. PhD Thesis. Rutgers University (October 2009)
14. Zheng, L., Cox, I.J.: Entropy-Based Static Index Pruning. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 713–718. Springer, Heidelberg (2009)
15. Zou, F., Wang, F.L., Deng, X., Han, S., Wang, L.S.: Automatic Construction of Chinese Stop Word List. In: Proceedings of the 5th WSEAS International Conference on Applied Computer Science (2006)