Interactive Construction of Visual Concepts for Image Annotation and Retrieval

Hugo Jair Escalante
Universidad Autónoma de Nuevo León,
Graduate Program in Systems Engineering,
San Nicolás de los Garza, NL 66450, México
hugo.jair@gmail.com,

J. Antonio Gonzalez-Pliego, Ruben Peralta-Gonzaga
Universidad Autónoma de Tlaxcala,
Unidad Académica Multidisciplinaria, Campus Calpulalpan,
Zaragoza No.1, Calpulalpan, Tlaxcala, 90200, Mexico
jagpliego222@hotmail.com, peraltagonzaga@gmail.com

Abstract

This paper describes a web-based interactive approach for building visual vocabularies in support of image annotation and retrieval. Under the proposed formulation a semantic concept is associated with a specific visual representation, thereby, explicitly modeling the *'semantic gap'*. Visual concepts obtained in this way can be used for image annotation/retrieval, object recognition and detection for both images and video and for building visual resources like visual dictionaries. A distinctive feature of our approach is that it can be used with any collection of images, without requiring the users to provide training data. Also, the proposed formulation allows prototype-based classification which can be useful for real-time applications. We performed preliminary experiments on image annotation using two different databases. Experimental results show that the proposed approach is a promising solution to bridge the semantic gap. However, some aspects of the proposed formulation still need to be improved.

1 Introduction

Nowadays the bag-of-words (BOW) paradigm is among the most popular approaches for facing several challenges in computer vision. Methods based on the BOW formulation have been proposed for image segmentation [10], image annotation/categorization [15, 3], image retrieval [14, 17], image filtering [16, 2], video sequence analysis [11] and object recognition [4, 7, 3, 1]. Under the classical BOW formulation, a set of visual words (textons, keyblocks, visterms) are defined/learnt and images are represented by histograms that account for the occurrence of visual words through the image. The set of visual words is referred to as the visual vocabulary (VV) or visual codebook.

Current approaches to VV construction concentrate their efforts on identifying useful visual descriptors and developing features that can effectively take into account spatial context and localization besides appearance [13]; whereas a few works have approached the problem of visual vocabulary learning [15]. Despite that very effective descriptors and features have been developed for building VVs, related methods are still limited in the sense that they fail in modeling the relationship between low-level features and high-level semantics, a challenge known in the computer vision argot as the *'semantic gap'*. Moreover,

Luis Enrique Sucar and Hugo Jair Escalante (eds.): AIAR2010: Proceedings of the 1st Automatic Image Annotation and Retrieval Workshop 2010. , volume 1, issue: 1, pp. 47-57

most of current methods still require of training sets of manually labeled images, at different levels (i.e. pixelwise, region-level or image-level), for defining VVs.

This paper presents a new approach for facing the problem of VV construction. The novelty lies in that instead of learning a model to link words to pictures we create a visual representation for each word; allowing us to reuse VVs and to efficiently apply BOWs-based methods. As obtaining manually labeled data is expensive, we take advantage of the large amounts of images available in the web. We consider a user-in-the-loop scenario, with the goal of obtaining VVs that are as accurate as possible, where the effort from the users is considerably reduced. A new approach for using VVs for image annotation is proposed, which is based on ideas from prototype classifiers; the benefit of this formation is that it allows us doing efficient visual concept detection and provides a more compact representation, we also study other forms of using VVs for image annotation. We present experimental results that show that our proposal is a promising research direction that attempts to bridge the semantic gap in a straightforward way. However, several aspects of our model still can be improved in order to effectively support computer vision tasks.

1.1 Building visual vocabularies

Given a set of images, the usual process for building VVs is as follows: (i) images are segmented, either manually or using automatic methods; (ii) predefined visual features are extracted from regions; (iii) extracted features from all regions are clustered, usually by using the k-means algorithm; (iv) from the clustering process, a set of visual words is defined, commonly, the centers of the clusters are considered as the set of visual words; (v) regions are associated to a single visual word: the most similar to it; (vi) images (respectively, regions) are represented by histograms that account for the occurrence of the different visual words. The entire process is depicted in Figure 1.

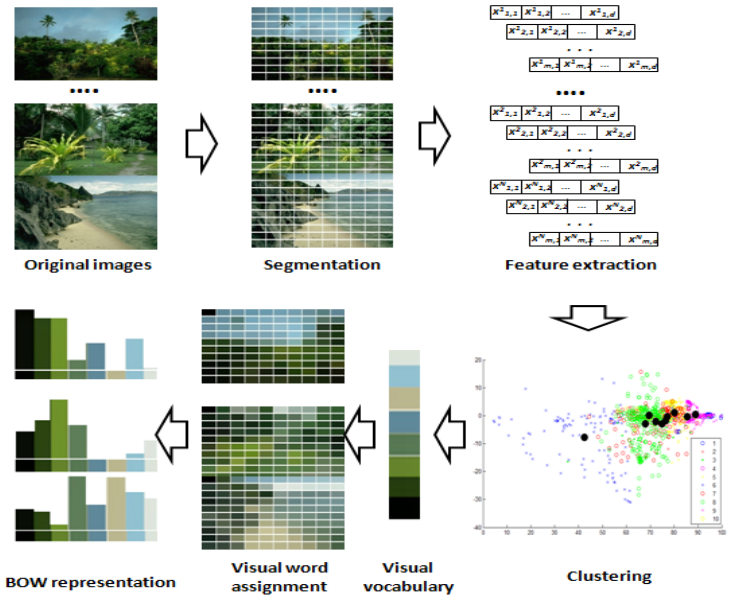

Figure 1: Illustration of the process of VVs construction. Images are segmented; features are extracted from each region; visual features are clustered; the centers of the clustering are considered visual words; images are represented by histograms that account for the occurrence of visual words in the image.

VVs provide a compact representation (the BOW representation) for images (respectively, regions), regardless of the dimensionality of the features considered in the extraction process. Furthermore, having discrete representations for visual information allows us to exploit powerful techniques from information

retrieval (e.g. the vector-space model and language models for information retrieval) and probabilistic modeling (e.g. probabilistic latent semantic analysis and latent Dirichlet allocation).

1.2 Using visual vocabularies

Depending on the application at hand, the VV and visual words are used in different ways. For example, for CBIR, the distances (e.g. Euclidean distance) between the BOW representation of documents in the image collection and that of a query image are calculated; documents are sorted in ascending order of their distance. Another option, introduced by Vogel et al., consists of using the BOW representation on a label basis; that is, first regions are annotated with semantic labels, hence images are represented by histograms of frequencies of the labels assigned to regions in images [14].

For object detection, images are represented by a combination of the BOW of regions in the image, these features are then used for training supervised learning classifiers (e.g. a support vector machine, SVM) [12, 13]; or for building probabilistic modeling classifiers that infer the association between words and pictures [2, 1].

The rest of this paper is organized as follows. The next section reviews related work on VVs construction. Section 3 describes the proposed methods. Section 4 reports experimental results on two image collections. Section 5 summarizes our findings and discusses future work directions.

2 Related work

Most techniques based on the BOW representation or using visual words in any form, construct the VV as depicted in Figure 1 [10, 3, 1, 12, 11, 7, 4, 17, 16, 2]. A notably exception is the work of Winn et al., where the authors propose a method for learning a universal visual vocabulary (UVD) for region labeling by adopting a Bayesian methodology; then they use the UVD with a Gaussian model and a k-nn classifier for labeling regions [15]. Their results using a k-nn classifier and the Gaussian model show the effectiveness of their method. The main contribution of Winn et al. is that they propose a principled way to learn visual features for region labeling, nevertheless, the learned features do not offer important advantages, in terms of region labeling accuracy, over using the usual codebook approach. Furthermore, this method requires of pixel-level annotated data, which is very expensive to obtain.

Joo-Hwee Lim proposes an interactive method for the construction of VVs for indexing and querying photographic collections [9]; under this formulation the VV is build manually: a user provides images containing the concept of interest, the user manually segments regions of interest and assign sub-labels to it. Hence, features extracted from both images and regions are considered to create the concept. Whereas the method can obtain effective visual concepts, it requires a significant effort from the user. Moreover, the user must provide the images to be used for creating the concept, delegating most of the process to the user.

Hentschel et al. proposed a method closely related to ours [8]. In their work they require the user to provide a set of images containing the concepts of interest. The user manually segments objects of interest in the images and assigns to each region the corresponding concept. Then the annotated segments are used to build a grid representation for smaller regions, around the ground truth segmentation. The resultant patches are clustered into $k-$groups and the centers of each group represent the visual words. Hence, $k-$visual words are associated to each label. The authors report experiments in an object recognition benchmark, although only per-class performance is reported; which compared to previous work on the same benchmark is rather limited. Another limitation of this approach is that manual segmentation and annotation is required, in this paper we use simple partitioning methods for solving this issue.

The main issue with both Hentschel et al.'s and Lim's approaches is that they require the user to

provide images containing each of the semantic concepts considered. In order to alleviate this issue, in this paper we use images publicly-available from the web for building the VVs. The web is the largest repository of information ever existed. Search engines allow users gathering multi-modal information to fulfill their information needs; because of this fact, several researchers have took advantage from information readily available in the web for enhancing computer vision applications [5]. Fergus et al. have proposed automatic techniques for filtering images returned by the ImageGoogle^{TM} search engine [6]. In a more recent work, the same authors have used images gathered from the web for training object recognition methods [5].

3 Interactive construction of visual concepts

We propose the construction of visual representations for specific concepts (labels, keywords) with the goal of supporting computer vision tasks. Under our formulation each concept is associated to a specific visual representation (i.e. a vector of features). The process we propose for building visual representations for concepts can be described using the scheme shown in Figure 1, with a few modifications, the difference lies in the way images are gathered and how the visual concepts are created. The specific process we consider is depicted in Figure 2.

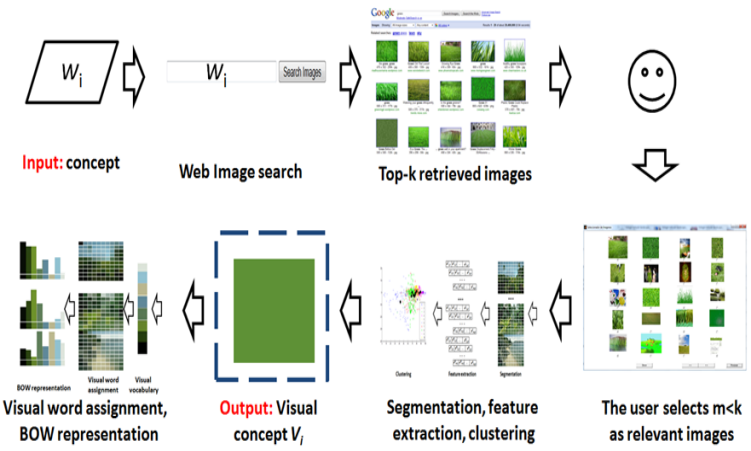

Figure 2: Diagram of the proposed methodology for interactive construction of VVs. An user provides a concept for which a visual representation is needed. $k-$ images are retrieved using ImageGoogle^{TM}. The user marks images relevant for the concept. A visual concept is obtained by using the marked images.

The user is given a list of concepts for which a visual representation is required. The user provides the list as input to the graphical user interface (GUI) we developed; for each concept, the GUI gathers $l-$images from the web, using the concept as query with the ImageGoogle^{TM} search engine; the GUI displays these $l-$images, together with selection buttons below each image. The user mark those images that are relevant for the concept at hand, we say an image is relevant if about 80% of the image is covered by the concept of interest, (the participation of the user ends here, thus as you can see her effort is very limited. Then, the marked images are used with the framework described in Section 1.1 (i.e. images are segmented into equal-sized patches, features are extracted from each region and regions are clustered together). Once image regions have been clustered we proceed to create the visual representation for a concept.

3.1 Creating visual concepts

For building visual concepts we consider two formulations. First, we followed a similar approach as that considered by Hentschel et al. [8], where, for each concept, all of the centers derived from the clustering method are considered visual words. Therefore, $k-$visual representations are associated with a single concept, that us, the representation for a concept c is a matrix $\mathbf{V}^c$ of dimensions $k \times d$, and where each row of $\mathbf{V}^c_{i,.}$ is u_i, the center of the i^{th} cluster, we call this formulation the *multi-word representation*.

Alternatively, we combine the $k-centers$ from the clustering process to obtain a single representation for each concept as follows

$$\mathbf{v}^c = \sum_{i=1}^{k} w_i^c \times u_i^c \tag{1}$$

where $\mathbf{v}^c \in \mathbb{R}^d$ represents the concept c; u_i^c is the center of cluster$-i$ from the clustering corresponding to regions of concept c; $w_i^c \in \mathbb{R}^1$ are scalars that weight the importance of the centers of each of the $k-$clusters for concept c, with $w_i^c \geq 0$ and $\sum_{i=1}^{k} w_i^c = 1$. Several options exist for defining w_i^c, including, the cohesion of each cluster and the entropy of samples associated to each cluster. For this work we considered a simpler approach, namely, the number of samples in the cluster. Thus we set:

$$w_i^c = \frac{|X_{u_i^c}|}{\sum_{j=1}^{k} |X_{u_j^c}|} \tag{2}$$

where $|X_{u_i^c}|$ is the number of regions associated to cluster i; intuitively, we will rely more on clusters that have associated a large number of examples to it, as such clusters will contain most of the regions relevant to the concept, whereas clusters with a few regions will contain regions that are noisy or barely associated with the concept. We call this formulation: *single-word representation*.

Note that the input to the system is a *'concept'*, which may lead to ambiguities, in principle we can use any label or set of labels in order to describe the concept we want to visualize. For example, if we want a signature for *'trunks'* in order to describe tree-like objects it would be useful to use: *'trunks trees nature landscape'*, this will lead to retrieve images related to natural landscapes as this is our goal. If instead we use the single label *'trunks'* we would mostly retrieve images of a cartoon character named *'trunks'*, among other irrelevant images (try *'trunks'* with ImageGoogle^{TM}); this example, also illustrates how to introduce prior knowledge into the visual concept construction. This freedom also form the basis for the development of visual signatures for concepts of higher semantics (e.g. we could build a concept for *'pornographic offensive images'* with the goal of filtering inappropriate images).

Once the visual representation for a concept is created we store it for its use on diverse tasks. In this paper we faced the problem of image annotation by using visual concepts as described in the next section.

3.2 Image annotation with visual concepts

We face the problem of image annotation at region-level, that is, we want to attach labels, from a predefined vocabulary, to regions in segmented images. We consider the following scenario: we are given the set of concepts that can be assigned to images $C = \{c_1, \ldots, c_M\}$; then we apply the methodology described in Section 3 for creating visual concepts for each of the $M-$concepts; we are also given a test set of images to be annotated, which are segmented and features are extracted from the resultant regions; then we use the set of visual concepts for assigning labels to regions.

For the multi-word representation we tried two variations. On the one hand, we consider the majority-vote strategy used by Hentschel et al. [8], where the distances between a test region $\mathbf{r}^T$ and the matrices for the visual concepts of all the words in the vocabulary are calculated. That is, we estimate distances

between the test region and every region in every matrix $\mathbf{V}_c$. We assign $\mathbf{r}^T$ the label that is mostly present in the $\kappa-$nearest neighbors of $\mathbf{r}^T$, we call this configuration *SAFIRE*.

Also, we consider the average distance of $\mathbf{r}^T$ to each matrix $\mathbf{V}_c$

$$D_{avg}(\mathbf{r}^T, \mathbf{V}^c) = \frac{1}{k}\sum_{i=1}^{k} D(\mathbf{r}^T, \mathbf{V}_i^c) \tag{3}$$

where $D(x,y)$ is the Euclidean distance between vectors x and y. Thus, we assign to $\mathbf{r}^T$ the label associated with the minimum average distance: $l^T = argmin_c D_{avg}(\mathbf{r}^T, \mathbf{V}^c)$, we call this setting *MWMAD*.

For the single-word representation we adopted a simpler methodology that consists in assigning a region $\mathbf{r}^T$ the label associated with the vector concept $\mathbf{v}_c$ for which the distance with $\mathbf{r}^T$ is minimum: $l^T = argmin_c D(\mathbf{r}^T, \mathbf{v}^c)$, we call this configuration *SWMD*.

Additionally, we build classifiers using the set of all regions corresponding to all of the concepts under consideration; that is, a training set for a multiclass classification problem. We consider different learning algorithms for building classifiers under this formulation, which we called *CLASSIFIER* (unless stated in the text we used the random forest classifier as learning algorithm), see Section 4.

There are a number of parameters that have to be tuned depending on the application and prior domain knowledge, these are: k the number of clusters to use, the size of the patches for image segmentation, the visual features to use, the web search engine to use and the assignment of words to obtain the BOW representation for images. In this paper we fixed all of these variables heuristically, although in future work we will explore the influence of these parameters into the performance of the methods that use the constructed visual concepts.

The GUI has been programmed in Matlab^R. For gathering web images we used a Perl implementation publicly available[1]. For clustering we used the $k-means$ algorithm, in particular we consider the implementation from the CLOP toolbox[2]; for experiments involving classifiers we used methods from this toolbox as well. We consider the following features in the current implementation: *average, standard deviation and skewness in RGB, CIE-Lab color spaces and color histogram in HSI.*

4 Experimental results

We report results of experiments on two data sets, described in Table 1; sample images from the considered data sets are shown in Figure 3. Each image from the VOGEL collection is segmented into 100 square regions, while the images in the MSRC data set have been manually segmented at a pixel level. For both collections labels have been assigned manually by humans.

Table 1: Data sets considered for our experiments.

Collection	Images	Regions	Labels	Reference
VOGEL	700	70,000	9	[14]
MSRC	591	2,062	21	[10]

We created visual concepts for the union of labels in both collections. Thus, the same visual concepts are used for labels appearing in both collections, the overlapping labels are: *grass, flower sky and water.* For the MSRC we report annotation accuracy in the test-set partition as provided by Shotton et al. [10].

[1] `http://search.cpan.org/~grousse/WWW-Google-Images-0.6.5/`
[2] `http://clopinet.com/CLOP`

Figure 3: Sample images from the MSRC (top row) and VOGEL (bottom row) collections.

For the VOGEL data set we annotated the 70,000 regions as described in [14]. Accuracy is measured as the percentage of regions that were annotated with the correct label.

Table 2 shows the accuracy obtained on region labeling of the different methods for building VVs as described in Section 3.2. In both data sets, the best result is obtained with the single-word representation (SWMD) for concepts, its accuracy is even higher than that of a random-forest (RF) CLASSIFIER using all of the collected data, thus giving evidence of the validity of our prototype-based technique for image annotation. Despite being low, we believe that the accuracy obtained by using visual concepts with the SWMD technique are competitive, since the human intervention required for building these data sets was of about 40 seconds per concept, which is much more smaller than that one would require for building training sets. Note that we are using heterogeneous images (from different contexts, with different illuminations, resolutions and scales) for labeling homogeneous images (i.e. images in each collection are similar to each other). Previous works have used the same type of images for training and testing, which makes the problem easier. It is important to emphasize that the results obtained with SWMD, SAFIRE and CLASSIFIER, outperform significantly the random annotation accuracy (11.11% for VOGEL and 4.76% for MSRC).

Table 2: Region labeling accuracy for the considered methods.

Method	**VOGEL**		**MSRC**	
	Accuracy	Time (s)	Accuracy	Time (s)
SWMD	**31.51**%	**0.18**	**15.94**%	**4.32**
MWMAD	4.39%	80.19	1.27%	0.49
SAFIRE(k=5)	20.79%	278.38	7.49%	1.54
KNN	22.38%	296.26	2.07%	1.23
CLASSIFIER	29,91%	1387.51	2.71%	1,101.23
Reference (Best)	71.70%	-	70.5%	-

The last row in Table 2, shows the accuracy obtained by reference methods that have used the same collections. Vogel et al. performed $10-$fold cross validation for achieving 71.7% of accuracy [14]; thus this result is the average over 10 trials of training-testing, where 63,000 and 7,000 regions are used for training and testing, respectively, with a support vector classifier. Therefore, one can roughly say that about 63,000 manually labeled regions are required for obtaining an accuracy of 71.7% on 7,000 regions; manually labeling 63,000 regions would take from weeks to months of hard work. Our method required

about 6 minutes of user interaction for building the visual concepts and a few seconds for annotating the $70,000$ regions. Thus we believe our view offers a good tradeoff between supervision needed and accuracy.

On the other hand, Shotton et al. proposed a conditional random field for region labeling, which is trained using more than $1,000$ regions manually segmented and labeled, they obtain 70.5% accuracy in the same partition we used. The accuracy with our methods is lower for the MSRC data set, this can be due to: 1) there are 21 classes for this data set which makes more difficult the annotation process; 2) regions in this data set have been segmented manually at pixel level, whereas the visual concepts were obtained from squared regions.

The smaller processing time is for the SWMD method, requiring of 18 seconds for the 70,000 regions in the VOGEL collection (i.e. 2.57×10^{-4} per image) and 14 seconds for the MSRC collection (i.e. 4.9×10^{-3} per image), showing the suitability of this method for real time applications.

Table 3 shows the confusion matrix for the SWMD method on the VOGEL data set. From this table we can see that there are several classes that are severely confused, these are: *flowers, rocks and grass*; also, *ground* is mostly confused with *sand* and *water* with *sky*, which is not surprising as both labels have very similar visual representations. The confusion can be due to the fact that the features we consider may not be appropriate for discriminating among those categories. In the future we will explore other types of visual features and we will incorporate further information that can help to avoid these mistakes (e.g. contextual and spatial information). The high confusion rates can also be due to the fact that images that were used for building one concept may also include other concepts (e.g., images for building the *flowers* concept often included grass).

Table 3: Confusion matrix for the SWMD method on the VOGEL data set. Rows show the truth labels and columns the predicted ones.

T/P	**G**	**S**	**W**	**F**	**T**	**S**	**G**	**R**	**FL**
Grass	**2.11**	20.55	43.30	18.78	3.14	6.62	5.48	0	0
Sky	0.13	**75.41**	14.80	1.29	0.13	7.78	0.38	0.01	0.03
Water	1.50	40.61	**30.98**	15.46	4.34	4.75	2.30	0	0.01
Foliage	12.80	9.00	22.80	**27.14**	18.44	4.04	5.56	0.08	0.09
Trunks	13.90	7.75	13.04	25.47	**27.38**	4.36	7.38	0.36	0.30
Sand	0	30.02	14.44	6.41	1.43	**41.71**	5.50	0.45	0
Ground	3.65	9.86	10.62	15.80	12.46	38.68	**8.83**	0.07	0
Rocks	1.29	12.83	18.10	26.62	17.93	15.97	7.19	**0.01**	0.01
Flowers	11.66	1.56	4.92	30.50	33.04	12.39	5.80	0.04	**0.04**

With the goal of evaluating the quality of the regions gathered from the web, we compared the annotation performance of using visual concepts created with images from the web to the performance of visual concepts built with hand labeled regions. For this experiment we consider the VOGEL data set. A subset of $L-$regions were randomly selected from the $70,000$ data set, we set L equal to the number of regions generated in the process of building visual concepts from the web, for our experiments we used $L = 9,664$. Then we used the selected regions for creating concepts as described in Section 3.2. Results of this experiment are shown in Table 4.

For most of the considered methods the difference is not significant, although the performance obtained by the random forest classifier is significantly improved when using hand-labeled regions, this accuracy is even comparable to that reported by Vogel et al. [14]. These results suggest that when manually labeled data are available the the combination of the visual words approach and the features we considered may not be the best way to characterize the labels. Also, as expected, the results indicate that it is desirable to use hand labeled data for supervised classification methods. Nevertheless, one should note that collecting such hand labeled data is expensive and time consuming, also, the SWMD is about

Table 4: Comparison of accuracy when building concepts from the web and when using ground-truth data

Method	**WEB-DATA**	**MAN-DATA**
SWMD	31.51%	32.25%
MWMAD	4.39%	4.38%
SAFIRE(k=5)	20.79%	20.76%
KNN	22.38%	14.04%
CLASSIFIER	29,91%	68.41%

7,700 times faster than the classifier we consider, see Table 2.

The results shown in Table 4, suggest that the collected data is not well suited for training supervised methods for classification. In order to confirm this result we compare the performance of several classifiers using the data gathered from the web, we consider those classifiers available in the CLOP machine learning toolbox. Results of this experiment are shown in Table 5. The highest performance is obtained

Table 5: Accuracy for different classifiers on the VOGEL data set.

Method	**Accuracy**	**Time**
Zarbi	29.34%	21.05
Naive	3.37%	26.12
Klogistic	25.55%	159.32
Neural	24.64%	351.16
Kridge	24.42%	329.30
Random forest	29.91%	1232.34

by the random forest classifier. With exception of the naïve Bayes classifier, the performance of all of the classifiers is comparable, note that the performance of the proposed SWMD method is higher than all the classifiers we considered.

Finally, Figure 4 shows the soft-labeling performance of SWMD for both web and ground-truth data. This plot shows the accuracy one would obtain if we look for the correct label in the $p-$more likely labels, sorted by its distance. We can see that the performance of the classifier built on ground-truth data is higher, achieving about 85% accuracy by considering $p = 2$ candidate labels. For the SWMD formulation one could obtain about 60% of accuracy by considering 3 candidate labels. This results motivate the development of postprocessing techniques that can refine the initial annotation.

5 Conclusions

We described an interactive approach for building visual codebooks in which no training images are required a priori and where the participation required from the user is small. Opposed to previous works on VV construction, our formulation associates a semantic concept with a visual representation. We proposed a prototype-based approach for image annotation and showed results on image annotation by using the generated concepts. Our results suggest the single-word representation offers potential performance-benefits and proved to be efficient, these advantages can be exploited for diverse computer vision tasks. Despite results with the visual concepts were positive, there are several aspects in which our methods can be improved, thus paving the way for further research on the subject. Future work directions include using more powerful visual features for representing the regions, using contextual and

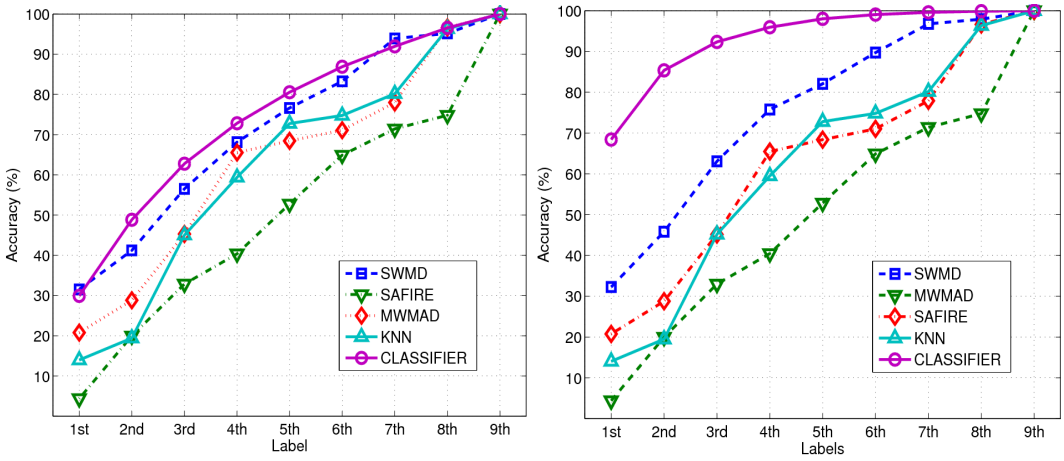

Figure 4: Soft labeling performance of the different methods using the web (left) and ground truth (right) data.

spatial information in the development of concepts, and the application of VVs for different tasks.

References

[1] K. Barnard, P. Duygulu, N. de Freitas, D. A. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3(1):1107–1135, 2003.

[2] T. Deselaers, L. Pimenidis, and H. Ney. Bag-of-visual-words models for adult image classification and filtering. In *Proc. of the Intl. Conference on Pattern Recognition*, pages 1–4, Tampa, Florida, USA, December 2008. IEEE.

[3] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of the European Conference on Computer Vision*, pages 97–112, London, UK, 2002. Springer.

[4] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531, San Diego, CA, USA, 2005. IEEE.

[5] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from googles image search. In *Proc. of the IEEE Intl. Conference on Computer Vision*, pages 1816–1823, Washington, DC, USA, 2005. IEEE.

[6] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *Proc. of the European Conference on Computer Vision*, volume 3021 of *LNCS*. Springer, 2004.

[7] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *Proc. of the Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, Alaska, 2008.

[8] C. Hentschel, S. Stober, A. Nürberger, and M. Detyniecki. Automatic image annotation using a visual dictionary based on reliable image segmentation. volume 4918 of *LNCS*, pages 45–56. Springer, 2007.

[9] J. H. Lim. Building visual vocabulary for image indexation and query formulation. *Pattern Analysis and Applications*, 4(1):125–139, 2001.

[10] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context. *IJCV*, 81(1):2–24, 2008.

[11] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *ICCV'03: Proc. of the 9th IEEE Intl. Conference on Computer Vision*, pages 1470–1477, Nice, France, 2003. IEEE.

[12] J. van Gemert, J. Geusebroek, C. Veenman, and A. Smeulders. Kernel codebboks for scene categorization. In *Proc. of the European Conference on Computer Vision*, pages 696–709, Marseille, France, 2008.

[13] K. van Sande, Theo Gevers, and C. Snoek. Evaluation of color descriptors for object and scene recognition. In *CVPR'08: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, Alaska, 2008.

[14] J. Vogel and B. Shiele. Semantic modeling of natural scenes for content-based image retrieval. *IJCV*, 72(2):133–157, 2007.

[15] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proc. of the Intl. Conference on Computer Vision*, Washington, DC, USA, 2005. IEEE.

[16] S. J. Yoo. Intelligent multimedia information retrieval for identifying and rating adult images. In *Proc. of the Intl. Conference on KES*, volume 3213 of *LNAI*, pages 164–170, Wellington, New Zealand, 2004. Springer.

[17] L. Zhu, A. B. Rao, and A. Zhang. Theory of keyblock-based image retrieval. *ACM Transactions on Information Systems*, 20(2):224–257, 2002.