Integrating visual words as bunch of n-grams for effective biomedical image classification

Glauco V. Pedrosa[1], Md Mahmudur Rahman[2], Sameer K. Antani[2], Dina Demner-Fushman[2], L. Rodney Long[2], Agma J. M. Traina[1]
[1]University of São Paulo - ICMC, São Carlos/Brazil
[2]National Institutes of Health - National Library of Medicine, Bethesda/USA
gpedrosa@icmc.usp.br

Abstract

The Bag-of-Visual-Words (BoVW) has been frequently used in the classification of image data. However, this modeling approach does not take into consideration the spatial relationships of these words, which is important for similarity measurement between images. We have developed a novel technique to incorporate spatial information of visual words based on the n-grams representation. The method encodes regional layout with a 2-gram representation in the local keypoint neighborhood. The region is divided in two zones to capture the relative orientations of pair-wise visual words. In turn, each image is described by an accumulated vector of 2-grams. Then, we compute the Shannon entropy over a random "bunch" of 2-grams to reduce the dimensionality of the feature vector. We discovered that this reduction technique creates a more discriminative feature vector as well as presents a considerable dimensionality reduction of up to 99%. The final representation is a compact and efficient local image descriptor that encodes frequency and arrangement of visual words. The proposed approach was tested by classifying a standard biomedical image dataset into categories defined by image modality and body part. The experimental results demonstrate the importance of contextual relations of visual words. Our proposed approach improved the classification accuracy compared to the traditional BoVW by 6.03%.

1. Introduction

Medical images have multiple modalities and are often captured with different views, imaging and lighting conditions. This results in a variety of image appearances and there has been high interest in developing robust invariant features for medical image description.

Image representation by local features is currently the state of the art in the field of computer vision area. The Bag-of-Visual-Words (or Bag-of-Keypoints or Bag-of-Features) approach has become an efficient method for modeling the local image features. This technique aims at describing an image as a histogram of *visual words*. Visual words are local image features, which concentrate relevant semantic information about the image. Each local feature is assigned to a visual word according to a pre-defined visual dictionary. The traditional representation is given by the frequency of each visual word contained in the image, similar to the Bag-of-Words approach in textual information. However, the Bag-of-Visual-Words model disregards all information about the spatial relationships of the features, affecting the accuracy of recognition.

The spatial information is critical for image and object characterization. Varying the image position/placement results in different features. Thus, the need of encoding the spatial relationship of visual words has motivated the creation of approaches to tackle this problem [5, 7]. Although the spatial information of visual words is important for visual characterization, the frequency of occurrence, which is captured by the "bag" or histogram, is also very important. Therefore, combining the frequency of occurrence and spatial information of visual words should be a promising direction for improving the image characterization. However, this direction has received relatively little attention. The main reason is computational, because modeling explicit spatial relationships among visual words is computationally expensive due to the large number of visual words in an image. It can considerably increase the feature vector dimensionality, which degrades the efficiency of processing similarity queries.

To overcome the weakness of the Bag-of-Visual-Words model, this paper introduces a novel method for encoding spatial information of local image features. The proposed method aims at providing an effective and efficient technique to extract *frequency* and *appearance* of visual

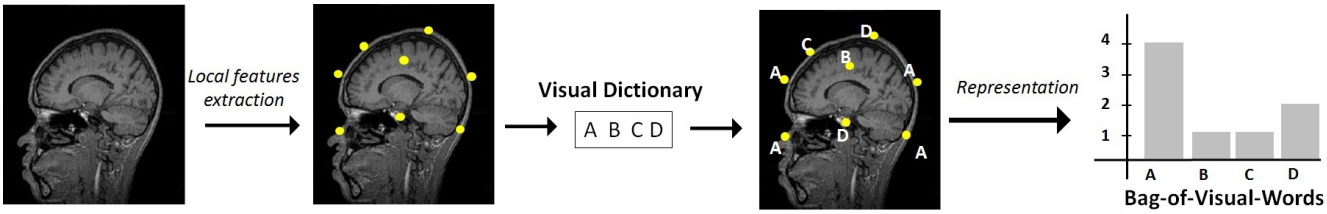


Figure 1. The Bag-of-Features approach represents the image as a collection of local features. For this, first the image local features are extracted, and then they are assigned to their nearest matching term according to a visual vocabulary (dictionary). The image feature vector is represented by the frequency of visual words detected in the image.

words within images considering their spatial location. From the linguistics field, we borrow the concept of n-grams and employ them for generating *visual phrases*. The n-gram is an efficient model widely used in natural language processing to represent textual documents. However, the use of n-grams is not a trivial solution to encode spatial arrangement of words, because the number of all possible combinations of n-grams increases exponentially with n. That is, given a dictionary with m words, the number of all possible n-grams is m^n. This challenges the use of n-grams as a modeling approach to encode spatial information of visual words, because we need to seek an effective image descriptor coupled with a compact representation. To tackle this problem, in this paper we focus on the use of 2-grams (bigrams). Our proposed method takes into consideration the frequency and orientation (angle) of pairs of visual words within a circular region around each *keypoint*. A keypoint is an "interest point" in the image, as determined by an automated method, such as SIFT. Then, we utilize an efficient method associating the Shannon entropy and *Bunch-of-2-grams* (see Section 3.3) to produce a new compact feature vector representation.

The proposed image representation encodes spatial information of visual words without the need of selecting features or the user interaction. We performed experiments using a public image dataset for biomedical image classification. The results demonstrate the importance of contextual relations of local features in the Bag-of-Visual-Words representation.

The remainder of this paper is structured as follows: Section 2 gives the formal definitions and the background needed to follow the work including the motivation. Section 3 explains the proposed method to encode spatial information of visual words. Section 4 presents the experimental analysis and Section 5 gives our conclusions.

2. The Bag-of-Visual-Words approach

The Bag-of-Visual-Words (BoVW) has emerged as an effective modeling approach to represent local image features. These local features are described by an *unordered* set of keypoints, where each keypoint describes representative local image features. The goal is to quantize these features using a visual dictionary. Figure 1 outlines the BoVW approach.

The main idea of using visual dictionaries is to consider that the local image visual patterns are similar to textual words in textual documents. Therefore, an image is composed of visual words as a textual document is composed of textual words. The final image representation employing the BoVW model is given by the frequency of each visual word in the image, analogously to the Bag-of-Words approach used in textual representation.

Some works employ correlograms of visual words [10] and image splitting by linear and circular projections [1] to encode the occurrences of the visual words in relation to the other visual words positions. It seems plausible that grouping words might be applied successfully for enriching the BoVW representation [14], once a high semantic information level is reached. The benefits of using groups of words have been proven to boost local feature matching [13]. Previous works employ *phrases* to model the co-occurrences of the words in local neighborhoods, using methods to encode the spatial layouts with a grid-dependent size [4]. A difficulty of these approaches is that the number of phrases can exponentially grow according to the number of words in a phrase. Thus, it is necessary to select a subset from the entire phrase set. Sophisticated mining or learning algorithms have been proposed for this selection [12], but it may still be risky to discard a large portion of phrases, because some may be critical for characterizing the images.

In this paper, we propose a more compact and robust spatial characterization considering *bunches* of 2-grams without the need of selecting representative features. Our proposed method also uses a simple and different procedure to extract the 2-grams from the image compared to [6], [11] and [9]. The proposed 2-gram extractor takes into consideration the orientation of 2-grams with a rotation invariant approach. The details of the proposed technique is presented in the next section.

3. The proposed method

As previously stated, the proposed method aims at encoding spatial arrangement of visual words using the

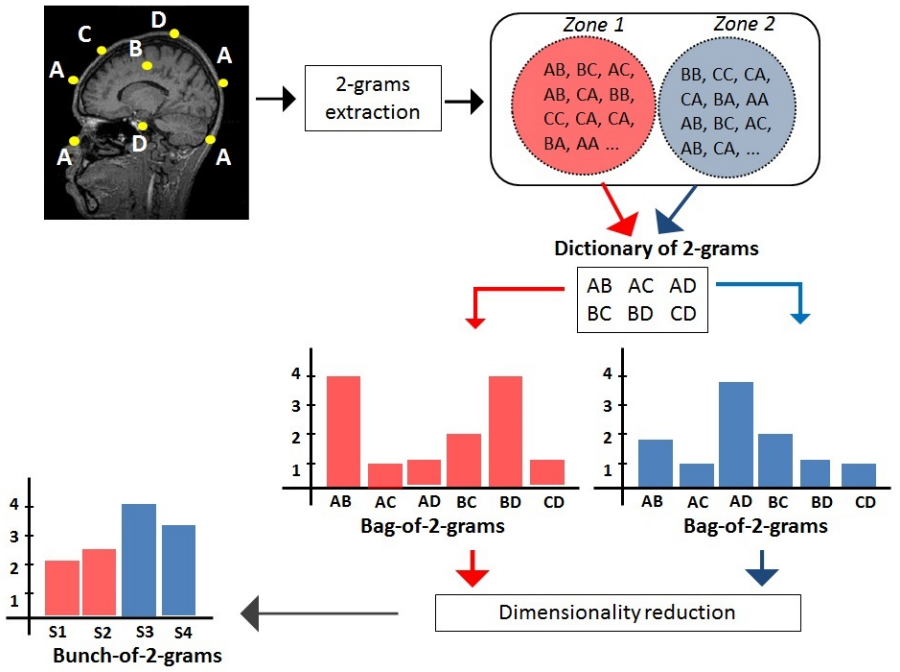


Figure 2. Proposed method to encode spatial information of visual words. See Figure 3 for zone definition

language model idea of n-grams. In our proposed model, the spatial information is the frequency and orientation of 2-grams extracted from the image. Figure 2 outlines the main steps of the proposed method. First, we extract the 2-grams from the image and put each of them in a specific bag according to its orientation. In this work we only consider two zones of orientation to encode the angle between pair-wise words. After that, the frequency of each 2-gram is computed according to a defined dictionary of 2-grams. Then, we group randomly sets of 2-grams and compute the Shannon entropy over these groups to create a new and compact feature vector. Our final image representation also takes into consideration the frequency of independent visual words in the image. In the following we explain in details how to extract the 2-grams from an image and how to represent an image using our proposed approach.

3.1. Extracting the 2-grams from an image

The visual 2-grams of an image can be generated by placing a region over each image keypoint. All pairs of words in this region formed with the center point are 2-grams.

We use a circular area of radius r around each image keypoint to extract 2-grams from an image. We divided this area into four 90° sectors as shown in Figure 3, then associate opposing sectors to create two zones (labeled red and blue in the figure).

Formally, let $P = \{p_1, p_2, ..., p_m\}$ be the local features detected in an image. Each local feature p_i is represented by its coordinates (x_i, y_i) in the image and it is assigned to a visual word w_j.

Definition 1. The Nearest Visual Words (NVW) of a local feature p_i is given by:

$$\text{NVW}(p_i, r) = P' \subset P \tag{1}$$

where

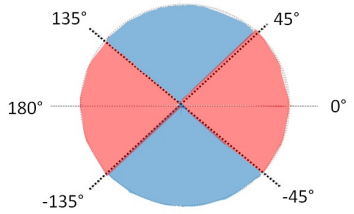
$$P' = \{p_j \in P' | \quad d(p_i, p_j) < r\} \tag{2}$$

and

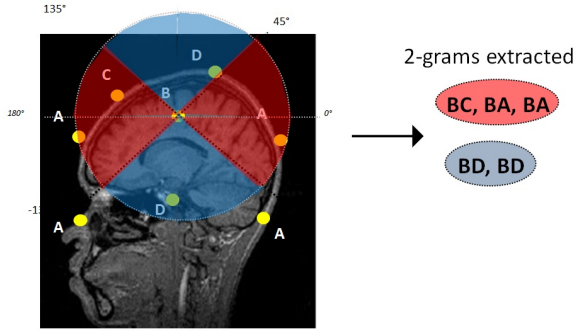
$$d(p_i, p_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{3}$$

Rotation invariance. In order to ensure rotation invariance, we use the information of the gradient direction calculated over a region around the keypoint. The initial angle of the circular area is placed over the computed gradient direction.

Figure 3(b) illustrates an example of 2-grams extracted from an image considering only one keypoint. This procedure is performed for each image keypoint. At the end, we will have two Bags-of-2-grams: one bag for 2-grams with angle within the interval [−135°,135°] and [−45°,45°], and other bag for 2-grams with angle within the interval [135°,45°] and [−135°,−45°]. Then, we compute the frequency of 2-grams for each bag according to a dictionary of 2-grams.



(a)



(b)

Figure 3. (a) Zones used to compute the 2-gram orientation, (b) 2-grams extracted considering a keypoint and the proposed circular area.

3.2. Bag-of-2-grams

After extracting the 2-grams from an image, each 2-gram is treated as a *visual phrase*. The next step is to create a Bag-of-2-grams by counting the frequency of *visual phrases* according to a dictionary of 2-grams. This dictionary could be all of the possible combinations of 2-grams. However, the size of this dictionary will be exponential with respect to the number of the visual words in the vocabulary. For example, considering 100 different visual words, the number of all possible combinations of 2-grams is 100^2.

To decrease the dictionary size, we (1) eliminate 2-grams consisting of repetitions of the same visual word (AA, BB, etc.) and (2) we identify a 2-gram with its corresponding inverted 2-gram (AB-BA, BC-CB, etc.), so that a 2-gram and its inverted form are counted as one visual phrase in the dictionary.

For example, the 2-grams generated by a dictionary formed by words $\{A, B, C, D\}$ are $\{AB, AC, AD, BC, BD, CD\}$. With the two restrictions proposed, the dictionary size is reduced from k^2 to $(k * (k-1))/2$, where k is the number of visual words. Even so, the dictionary size remains very large affecting the dimensionality of the feature vector. Taking into consideration that we have two bags, the dimensionality of the final image feature vector is given by:

$$k + 2 * ((k * (k-1))/2) \tag{4}$$

To reduce the dimensionality of the final feature vector, we develop a strategy to group sets of 2-grams for generating a more compact feature vector. As we will present in the following.

3.3. Bunch-of-2-grams

In this paper we take advantage of a method that reduces the dimensionality of the Bag-of-2-grams while retaining the discrimination power. Our approach takes advantage of the Shannon entropy measurement to achieve such reduction. The proposed idea is to compute the Shannon entropy over *bunches* of 2-grams. Entropy is defined as a measure of "uncertainty" or "randomness" of a random phenomenon. Suppose that some information about a random variable is received. Then a quantity of uncertainty is reduced, and this reduction in uncertainty can be regarded as the quantity of transmitted information.

Definition 2. Let $B = \{b_1, b_2, ...b_m\}$ the Bag-of-2-grams extracted from an image, where b_j is the frequency of occurrence of a specific 2-gram in the image. A *Bunch-of-2-grams* is given by a partition of B into t distinct groups of size q, this means, $B_t = \{C_1, C_2, ..., C_t\}$, so that $C_1 \cup C_2 \cup ... \cup C_t = B$, $C_i \cap C_j = \emptyset$ and $|C_i| = q$ for $i = 1...t$. Then, to produce a more compact feature vector we compute the Shannon entropy over each bunch as defined:

$$H(C_i) \equiv H(c_1, ..., c_q) \equiv -\sum_{i=1}^{q} c_i log c_i \tag{5}$$

where each small c_i is in bunch C_i.

To illustrate this concept let us consider a Bag-of-2-grams with $m = 500$ values, where each value corresponds to the frequency of occurrence of a specific 2-grams. If we divide this bag into 100 bunches with $q = 5$ elements, then we will have a final feature vector with 100 values representing the entropy of each bunch. This provides a considerable dimensionality reduction from 500 to 100. Modifying the quantity of elements q in each bunch to 10, 50 or 100, it will result in feature vector with 50, 10 and 5 values respectively. That is, the size of the reduced feature vector obtained when employing the proposed method is given by:

$$\text{vector size} = \frac{m}{q} \tag{6}$$

where, m is the size of the dictionary of 2-grams; $0 < q < m$ is the quantity of elements in each bunch, which must always be a power of the dictionary size.

Thus, each element of the resulting final feature vector contains the accumulated Shannon entropy calculated over

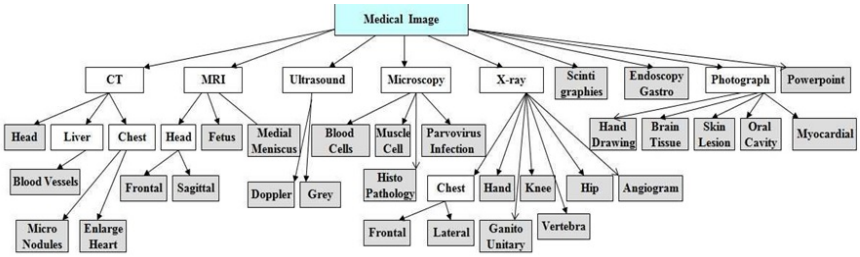


Figure 4. Classification structure of the medical image data set.

a bunch of 2-grams. The quantity of elements in each bunch is the dimensionality reduction factor for the original Bag-of-2-grams. Therefore the size of the new feature vector is equal to the number of bunches.

A weakness of this proposed dimensionality reduction is to choose the best partition of the Bag-of-2-grams. Different partitions could result in different classification accuracies. However, as we will show in the experiments section, a random partition yields a large gain in classification accuracy, when compared with the original Bag-of-2-grams.

4. Experimental Results

In this section, we evaluate the proposed algorithm for biomedical image classification. We use SIFT descriptors of 16x16 pixel patches computed over keypoints. We do multi-class classification with a support vector machine (SVM) trained using the one-versus-all rule: a classifier is learned to separate each class from the rest, and a test image is assigned the label of the classifier with the highest response.

The image dataset employed is composed of 5042 biomedical images of 32 manually assigned disjoint global categories, which is a subset of a larger collection of six different data sets used for the medical image retrieval task in ImageCLEFmed 2007 [8]. In this collection, images are classified into three levels as shown in Fig 4. In the first level, images are categorized according to the imaging modalities (X-ray, CT, MRI, etc.). At the next level, each of the modalities is further classified according to the examined body parts (head, chest, etc.) and finally it is further classified by orientation (frontal, sagittal, etc.) or distinct visual observation (e.g. CT liver images with large blood vessels). The 32 disjoint categories are selected only from the leaf nodes (grey in color) to create the ground-truth data set. In each of our experiments we tested the accuracy of our algorithm in assigning one of the 32 category labels to an image.

We first tested the baseline BoVW using different dictionary sizes. The visual dictionary was generated by applying a clustering technique over keypoints detected by SIFT in a subset of images sampled from the dataset. We achieved good results using a dictionary of 450 visual words (clusters). The cross-validation accuracy achieved using BoVW and 450 words was 82.53%.

A dictionary with 450 words resulted in a Bag-of-2-grams with 101025 values. As we are using two bags, the final feature vector is given by 202500 values (see eq. 4). The dimensionality of this feature vector impairs its comparison and indexing, so we need to reduce its size to perform a classification test.

We tested the proposed reduction technique (section 3.3) using different bunch sizes to find the size with the best accuracy result. Table 1 summarizes the accuracy achieved for each bunch size tested. The best accuracy classification was achieved using the proposed technique with 449 bunches. This produced a final feature vector with 1,348 values. This result improved the accuracy of the traditional BoVW representation in 6.03%.

Table 1. Classification accuracies obtained using different bunch sizes.

Accuracy	Bunch size (q)	Final vector size
87.50%	2,245	4940
88.44%	1,347	3144
88.56%	**449**	1384
87.40%	225	900
85.26%	45	540
83.68%	3	456

We also performed several different random partitions to produce a *Bunch-of-2-grams* with 449 bunches. In general, the accuracies achieved were almost the same. The best accuracy result achieved was 89.2 % and the worst 88.1%.

Using a dictionary with 450 words and a bunch size of 449, we performed a test validation using 80-20 classification: 80% of the image dataset was used to train our classifier and 20% to test. Table 2 presents the results comparing the proposed technique with three other methods from the literature. Our proposed method achieved a performance of 10.13% superior compared to the second best result (CEED) and it improved the BoVW technique by 6.74%.

Table 2. Comparative results using 80-20 classification test.

CLD [2]	EHD [2]	CEED [3]	BoVW	Our Method
58.11%	58.94%	68.54%	71.93%	**78.67%**

5. Conclusion

In this paper, we have introduced a novel modeling approach for representing images by taking into consideration the spatial relationships among its visual words. The proposed method is based on a collection of *visual phrases*, instead of considering the image as a set of isolated visual words. Our proposed method is an analogy to the popular n-gram representation used for textual representation. In this work, we have focused specifically on the use of 2-grams for image representation due to its computational simplicity.

The proposed technique uses information about the angle between pair-wise words to encode orientation of the 2-grams. In this work we discretize the angles into two different zones. The 2-grams are extracted from the image and separated according to their zones. Then, we compute a Bag-of-2-grams for each zone. Finally, we reduce the dimensionality of the Bag-of-2-grams using the Shannon entropy calculated over a random combination of groups of 2-grams. By experimental results, we demonstrated that our proposed reduction technique considerably reduces the dimensionality of the feature vectors up to 99% and provides a significant gain in classification accuracy. An important characteristic of our method is that it does not require the selection of representative features or user interaction.

We evaluated our method by applying it for classifying biomedical images into categories determined by image modality and body part. Experimental results show that the proposed contextual relations between visual words is a powerful asset for improving the image categorization using the Bag-of-Visual-Words approach.

Acknowledgements

This research was supported [in part] by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

The authors also acknowledge the financial support of FAPESP, CNPq, CAPES, INCT INCod and SticAmsud.

References

[1] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial-bag-of-features. In *CVPR'10*, pages 3352–3359, 2010. 2

[2] S. F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 standard. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):688–695, June 2001. 6

[3] S. Chatzichristofis and Y. Boutalis. CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval. pages 312–322. 2008. 6

[4] Y. Jiang, J. Meng, and J. Yuan. Grid-based local feature bundling for efficient object search and localization. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 113–116, 2011. 2

[5] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society. 1

[6] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua. Contextual bag-of-words for visual categorization. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(4):381–392, 2011. 2

[7] J. Liu and M. Shah. Scene modeling using co-clustering. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–7, 2007. 1

[8] H. Mller, T. Deselaers, T. M. Deserno, J. Kalpathy-Cramer, E. Kim, and W. Hersh. Overview of the imageclefmed 2007 medical retrieval and medical annotation tasks. In C. Peters, V. Jijkoun, T. Mandl, H. Mller, D. W. Oard, A. Peas, V. Petras, and D. Santos, editors, *CLEF*, volume 5152 of *Lecture Notes in Computer Science*, pages 472–491. Springer, 2007. 5

[9] G. V. Pedrosa and A. J. M. Traina. From bag-of-visual-words to bag-of-visual-phrases using n-grams. In *SIBGRAPI 2013 (XXV Conference on Graphics, Patterns and Images)*, August 2013. 2

[10] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *In IEEE Computer Vision and Pattern Recognition*, pages 2033–2040, 2006. 2

[11] P. Tirilly, V. Claveau, and P. Gros. Language modeling for bag-of-visual words image categorization. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, CIVR '08, pages 249–258, New York, NY, USA, 2008. ACM. 2

[12] L. Torresani, M. Szummer, and A. W. Fitzgibbon. Learning query-dependent prefilters for scalable image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '09, pages 2615–2622, 2009. 2

[13] Y. Zhang, Z. Jia, and T. Chen. Image retrieval with geometry-preserving visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 809–816. IEEE, 2011. 2

[14] L. C. Zitnick, J. Sun, R. Szeliski, and S. Winder. Object instance recognition using triplets of feature symbols. In *Tech. Report, Microsoft Research*, 2007. 2