Creating Compact and Discriminative Visual Vocabularies Using Visual Bits

T. Kirishanthy and A. Ramanan
Department of Computer Science, University of Jaffna, Sri Lanka.

Introduction

The generic framework of a bag-of-features (BoF) approach is depicted in Figure 1 (1,8-15). The problem with this approach lies in that constructing a vocabulary for each single dataset is not efficient. Usually the construction of a vocabulary is achieved by cluster analysis.

- A larger size of vocabulary increases the computational needs.
- On the other hand, a smaller size vocabulary degrades the classification rate.

Therefore, the choice of the size of a vocabulary should be balanced between the recognition rate and computational needs.

Objectives

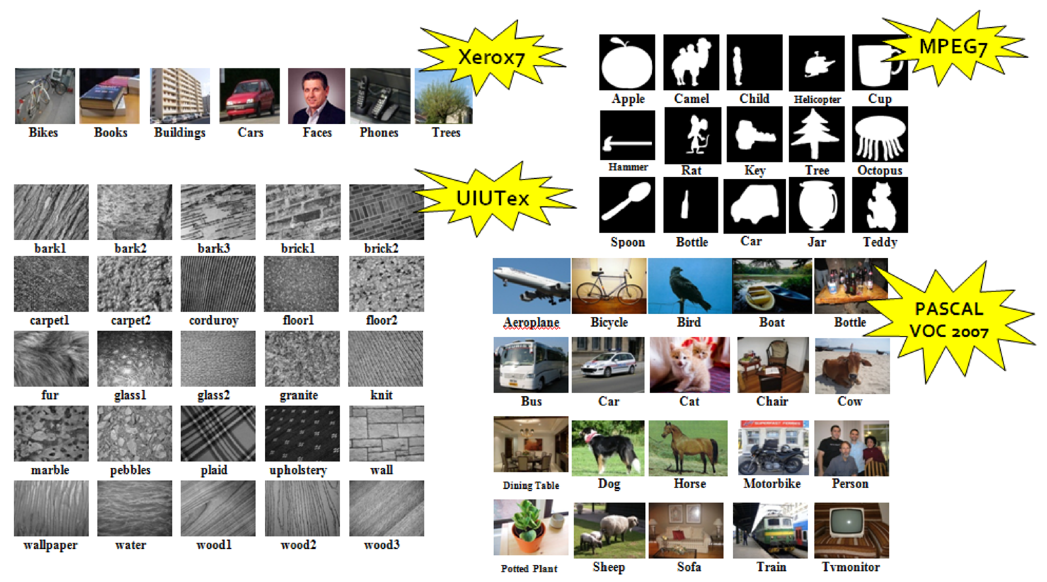
To map an initial high-dimensional vocabulary into a compact vocabulary while maintaining its discriminative power.

Methodology

The proposed vocabulary compression technique is depicted in Figure 1 (1-7,9-15). The proposed method maps an initial high dimensional visual vocabulary into a more compact form while maintaining its discriminative power. The reduction of BoF vocabularies to improve coding efficiency is achieved by two-step process:

1. Encode each image as "bits", i.e., the significant presence or absence of each visual word.
2. Remove visual words with bits that are not activated enough in images.

Experimental Setup



Xerox7: 7 classes, 1776 images [1]; **UIUTex**: 25 classes, 40 images/class [4]; **MPEG7**: 15 classes, 20 images/class [3]; **PASCAL VOC 2007**: 20 classes, 9963 images [2].

- Xerox7 & UIUTex: 70% training, 30% testing.
- MPEG7: 50%-50% training-testing.
- PASCAL VOC2007: Provided training & testing sets.
- Features: SIFT descriptors [5].
- Vocabulary Construction: K-means algorithm.
- Classification: Linear OVA-SVMs.

Methodology ...

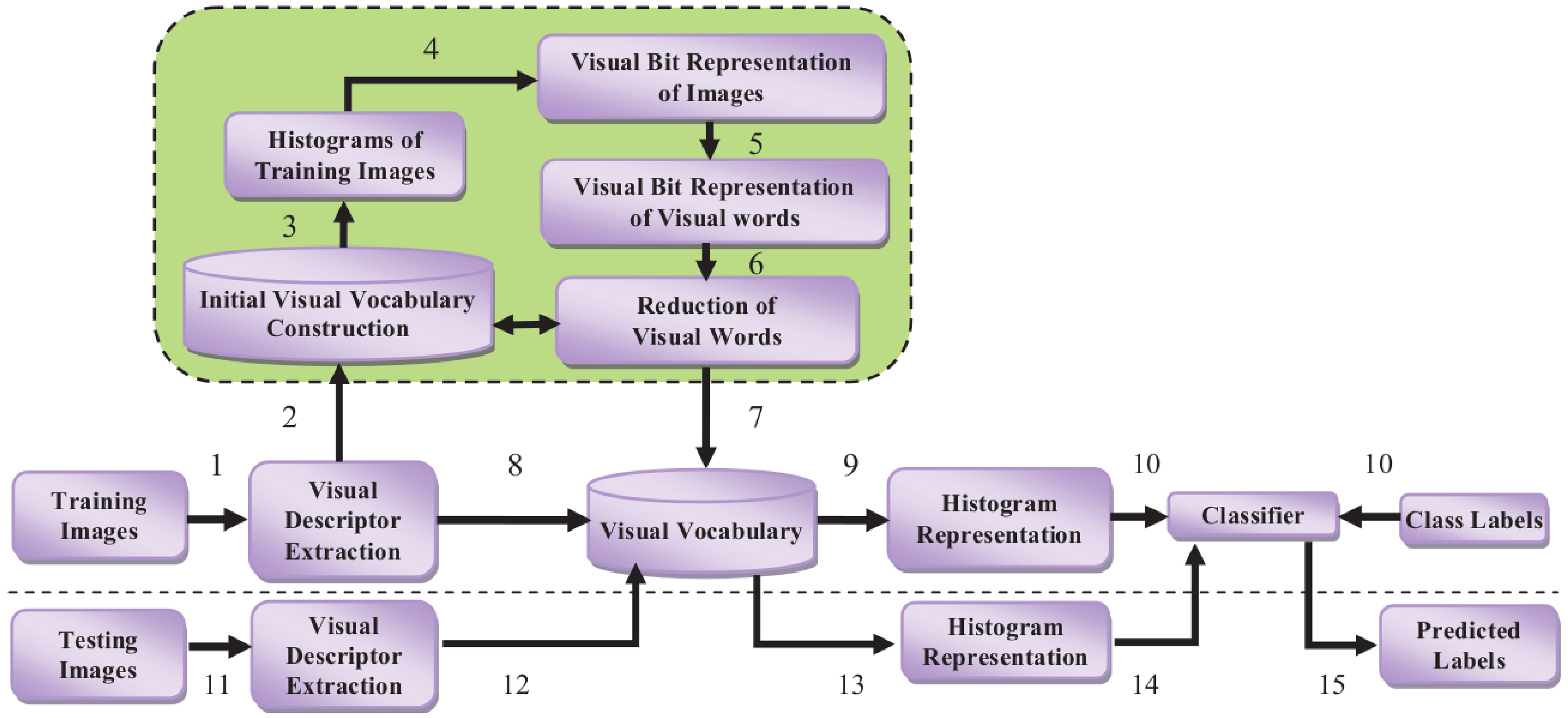


Fig. 1. Framework of creating a compact and discriminative visual vocabulary using visual bit representation. Below the shaded block, diagram in dotted outline shows the traditional bag-of-features (BoF) approach. The proposed method is shown in shaded block diagram (2-7) which adds an additional layer of compression to the traditional way of constructing a vocabulary in the BoF framework (1, 8-15).

Visual Bit Representation of Images

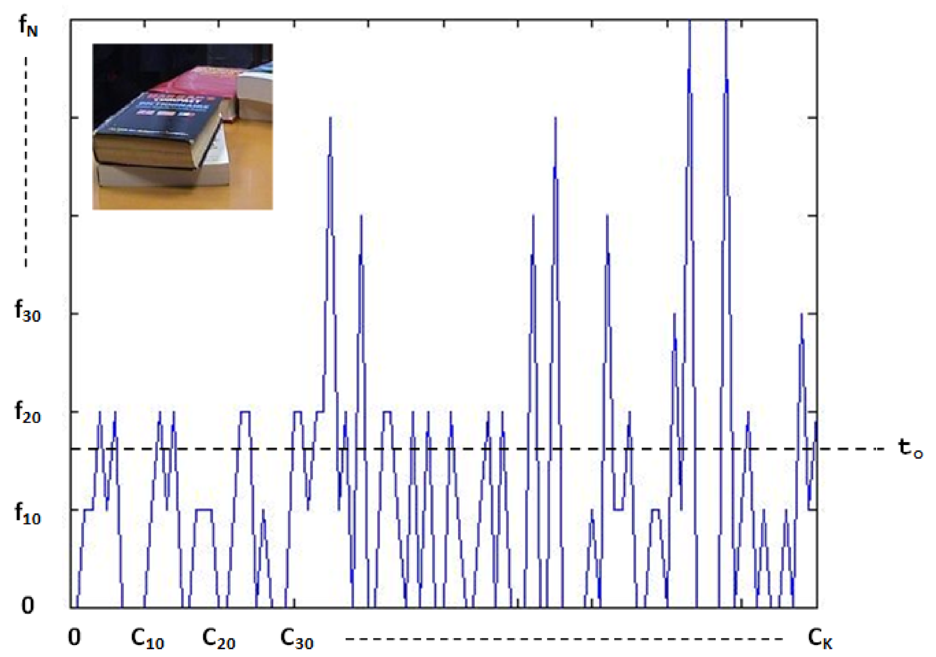


Fig. 2. Visual bit representation of images.

The patch-based descriptors of image, I, are mapped into a feature vector by computing the frequency histogram, h, with the initial vocabulary V. The average number of descriptors that fall into each visual word C_i of V is computed as t_0 for each image I of the training set. The visual bit representation of an image is then coded using equation (1).

$$h_i = \begin{cases} 1 : \text{if } (|C_i| \geq t_0) \\ 0 : \text{otherwise} \end{cases} \quad \forall i = 1, \ldots, K \tag{1}$$

where K is the size of initial vocabulary

This process is repeated to all training images of a specific-category by computing t_0 corresponding to an image.

Visual Bit Representation of Visual Words

	C_1	C_2	C_3	- - -	C_K
I_1	1	0	0	- - -	1
I_2	0	1	0	- - -	0
I_3	1	1	1	- - -	0
⋮	⋮	⋮	⋮		⋮
I_M	0	1	1	- - -	1
Total	SB_1	SB_2	SB_3	- - -	SB_K

Fig. 3. Visual bit representation of visual words.

Following the visual bit representation of images the initial vocabulary V is coded as a sparse representation by using equation (2) where SB_i indicates the sum of visual bits associated with the ith visual word.

$$t_1 = \frac{\lambda p_0 + p_1}{\lambda + 1} \tag{2}$$

where $p_0 = min_{1 \leq i \leq K}(SB_i)$ and $p_1 = max_{1 \leq i \leq K}(SB_i)$.

We can now compress the initial visual vocabulary using the subsequent step indicated by equation (3).

References

[1] C. Csurka, R. Dance, L. Fan, J. Willamowski, and C. Bray., "Visual Categorization with Bags of Keypoints", In Workshop on Statistical Learning in Computer Vision, (*ECCV*), pp. 1-22, 2004.

[2] M. Everingham, L. Van-Gool, C. K. I. Williams, J. Winn, and A. Zisserman., "The PASCAL Visual Object Classes Challenge 2007 Results", http://www.pascalnetwork.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[3] L. Jan Latecki, R. Lakamper and U. Eckhardt, "Shape Descriptors for Non-rigid Shapes with a Single Closed Contour", In proceedings of IEEE Conference on Computer Vision and Pattern Recognition (*CVPR*), pp. 424-429, 2000.

[4] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions", In IEEE Transactions on Pattern Analysis and Machine Intelligence (*PAMI*), pp. 1265-1278, 2005.

[5] D. Lowe., "Distinctive Image Features from Scale-invariant Keypoints", In International Journal of Computer Vision (*IJCV*), vol. 60, pp. 91-110, 2004.

Reduction of Visual Words

We learn the importance of each visual word of the initial visual vocabulary V through the visual bit representation of visual words.

$$\text{Compact}_{\text{CB}} = \begin{cases} \text{eliminate } C_i : \text{if } (SB_i < t_1) \\ \text{retain } C_i \quad : \text{otherwise} \end{cases} \tag{3}$$

where t_1 indicates the level of significant activation of a visual word in a category-specific vocabulary.

The same process described in reducing a category-specific vocabulary could also be applied to constructing a global vocabulary.

Testing Results

Table 1: Mean Classification rate with standard BoF approach having Category-specific vocabularies.

Dataset	Initial vocabulary (Traditional BoF)			Compact vocabulary (Ours)			
	K	NN	SVM	λ	size	NN	SVM
Xerox7	700	73.55	94.93	1	251	78.05	94.24
				2	459	76.55	95.18
				3	553	75.05	94.80
PASCAL 2007	1000	14.53	94.99	1	437	15.17	94.99
				2	668	15.88	95.00
				3	769	15.44	94.99
UIUCTex	1000	96.67	99.79	1	477	97.00	99.72
				2	634	96.00	99.77
				3	730	96.33	99.83
MPEG7PartB	600	44.67	97.56	1	198	58.67	97.42
				2	297	52.67	97.38
				3	373	51.33	97.29

Table 2: Mean Classification rate with standard BoF approach having Globally constructed vocabulary.

Dataset	Initial vocabulary (Traditional BoF)			Compact vocabulary (Ours)			
	K	NN	SVM	λ	size	NN	SVM
Xerox7	1000	71.67	94.56	1	248	75.61	94.48
				2	844	72.42	94.43
				3	957	72.23	94.75
PASCAL 2007	1000	13.49	94.98	1	142	13.90	95.00
				2	677	14.73	94.99
				3	909	13.76	94.99
UIUCTex	1000	95.33	99.69	1	785	96.00	99.65
				2	932	96.00	99.71
				3	958	95.00	99.68
MPEG7PartB	600	32.00	97.33	1	15	66.00	93.42
				2	101	58.00	96.31
				3	199	49.33	96.89

Discussion and Conclusion

- The proposed method yields compact vocabulary while maintaining its discriminative power.
- It provides a way to choose optimal vocabularies for recognition.
- The classification performance is comparable to or even better than the standard BoF approach.
- Needs less computational overhead.
- Guides the future works in BoF approach to deal with very low-dimensional representation.