Understanding Visual Dictionaries via Maximum Mutual Information Curves

Wei Zhang
Oregon State University
zhangwe@eecs.oregonstate.edu

Hongli Deng
ObjectVideo
hdeng@objectvideo.com

Abstract

Visual dictionaries have been successfully applied to "bags-of-points" image representations for generic object recognition. Usually the choice of low-level interest region detector and region descriptor (channel) has significant impact on the performance of visual dictionaries. In this paper, we propose a discriminative evaluation method -- Maximum Mutual Information (MMI) curves to analyze the properties of the visual dictionaries built from different channels. Experimental results on benchmark datasets show that MMI curves can give us not only insight into the discriminative characteristics of the visual dictionaries, but also provide straightforward guidelines for the design of the image classifier.

1. Introduction

Real-world object recognition datasets usually contain significant appearance variation in images due to different pose, visual transformation, occlusion, noise signals and so on. In order to obtain relatively invariant and compact representations of objects, various interest region detectors [1,2,3,7] have been applied to images to extract distinct and salient regions. Region descriptors are then commonly computed to describe the image contents within the interest regions. The most famous one is the SIFT descriptor [3].

Recently years have seen great success of visual dictionary [4,5] approaches for generic object recognition based on these local region descriptors. These approaches use clusters of region descriptors as the initial entries in the visual dictionary. The recognition task is then accomplished by manipulating the entries and selecting the most discriminative ones to build the final image classifier. Different combinations (channels) of interest region detectors and region descriptors will produce different pools of features to build the dictionaries. Ideally, all the entries should be consistent and informative to make classification a trivial task. The choice of detector has significant impact on the performance of recognition approaches [4,5]. But it is not always obvious which detector preferred for a given problem. Usually, the choice is made purely empirically. Given an object recognition problem, it would be much more rational to experiment only with detectors that are promising for the problem, rather than trying every available detector.

Different evaluation criteria [4,5] have been proposed to measure the discriminative ability of descriptor clusters. In [4], the discriminative power of the clusters is evaluated using the classification likelihood and mutual information criteria. In [5], the clusters of region descriptors are evaluated based on their average cluster precision. Motivated by previous work and the successful discriminative feature selection algorithm in [6], we propose the Maximum Mutual Information (MMI) evaluation criterion, which measures the discriminative power of visual dictionary entries quantitatively. Our evaluation method is closely related to the classification of image instances in the recognition task. It can be performed on any object recognition dataset efficiently without the requirement for prior knowledge of homographies. The MMI curves can clearly reveal the characteristics of dictionaries for the specific object recognition problem. Additionally, comparison results are valuable guidelines for the design of the image classifier. In this paper, visual dictionaries built from state-of-art interest region detectors are evaluated on benchmark datasets. The results can help future researchers to select suitable detectors for similar object recognition problems.

2. MMI evaluation method

2.1. Clusters learned by GMM-EM algorithm

Given a binary object recognition dataset composed of object (positive) images and background (negative) images, the positive images are partitioned into two disjoint sets, one is called the *clustering* set, denoted as I_C. The other positive set is combined with all negative images to form the *evaluation* set I_E. Then a specific interest region detector is applied to all the images. For each detected region, a SIFT descriptor [3] is computed to produce the *clustering* descriptor vectors F_C and *evaluation* descriptor vectors F_E.

As in [4], our method first fits a Gaussian Mixture Model (GMM) to the *clustering* descriptor vectors. Each cluster C_k is described by a d-dimensional mean vector $\boldsymbol{\mu}_k$ and a $d\times d$ diagonal covariance matrix $\boldsymbol{\Sigma}_k$. In our experiments, the number of clusters K is set to 50.

2.2. MMI score

Given: a cluster C_k: $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$; the *evaluation* set I_E contains I images; the class labels of evaluation images $L_E = (l_1, \ldots, l_i, \ldots, l_I)$, with $l_i \in \{+1,-1\}$; and the SIFT vectors of evaluation images $F_E = (F_1, \ldots, F_i, \ldots, F_I)$, we would like to evaluate the discrimination property of cluster C_k, that is, how well does the cluster reveal the categories of evaluation images. This is done by employing cluster C_k to classify the evaluation images, and search for the maximum mutual information (MMI) between the classification results and the true class labels, then take MMI as the evaluation score for C_k. In order to classify the evaluation images using cluster C_k, first, we calculate the distance from C_k to all the evaluation images. For cluster C_k: $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and image I_i, the distance between them is:

$$d_{k,i} = \min(d(F_i, C_k)) = \min_j((\boldsymbol{v}_{ij} - \boldsymbol{\mu}_k)^t \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{v}_{ij} - \boldsymbol{\mu}_k)) \quad (1)$$

where $\boldsymbol{v}_{ij} \in F_i$ is the SIFT descriptor vector computed from region $j = 1, \ldots, D_i$. D_i is the total number of detected regions in image I_i.

Given the distances between C_k and evaluation images: $(d_{k,1}, \ldots, d_{k,i}, \ldots, d_{k,I})$; all the evaluation images can be sorted based on the distances. That is, find a permutation $\boldsymbol{\pi} = (\pi(1), \ldots, \pi(i), \ldots, \pi(I))$ such that:

$$d_{k,\pi(1)} \leq \ldots \leq d_{k,\pi(i)} \leq \ldots \leq d_{k,\pi(I)} \quad (2)$$

Then the sorted class labels are:

$$\boldsymbol{l}_{k,\boldsymbol{\pi}} = (l_{k,\pi(1)}, \ldots, l_{k,\pi(i)}, \ldots, l_{k,\pi(I)}),\ l_{k,\pi(i)} \in \{+1,-1\} \quad (3)$$

The sorted label array illustrates the discrimination power of cluster C_k. A perfect cluster should have all the positive images (+1) ranked first followed by all the negative images (−1); while a poor cluster can never discriminate between them, so gives randomly ordered labels. From the view of Information Theory, the sorted label array $\boldsymbol{l}_{k,\boldsymbol{\pi}}$ indicates how much information the distance-based sorting can tell about class labels, which can be quantitatively measured by the maximum mutual information (MMI) between the sorting array and true class labels:

$$MMI_k = \max_s (MI(\boldsymbol{l}_{k,\boldsymbol{\pi}}, s)) \quad (4)$$

$MI(\boldsymbol{l}_{k,\boldsymbol{\pi}}, s)$ calculates the mutual information between the classification results $C_{k,s}$ and true image class L_E. $C_{k,s}$ is given by a decision stump [6] which set threshold at the position s ($1 \leq s \leq I$) in the sorted label array, and classifies the images before s to be positives, those after s to be negatives. The search for the maximum in (4) can be sped up by addressing the fact that the maximum can only possibly be obtained between +1 followed by a −1 in the sorted label array. Using the sorted label array $\boldsymbol{l}_{k,\boldsymbol{\pi}}$, the mutual information in (4) can be calculated similarly as in [4]. MMI scores measure the discriminative power of dictionary entries. A perfectly discriminative entry is assigned full MMI score of 1.0; while a non-discriminative entry will have a score near 0.

2.3. MMI curves

We evaluate the performance of visual dictionaries built from several state-of-art interest region detectors: (1) Harris Laplace detector (*HarLap*), Hessian Laplace detector (*HesLap*), and their affine-invariant versions, Harris affine-invariant detector (*HarAff*) and Hessian affine-invariant detector (*HesAff*) in [1]; (2) Difference-of-Gaussian (*DoG*) detector [3]; (3) Maximally Stable Extremal Regions (*MSER*) [7]; (4) Curvilinear Regions (*Curvilinear*) detector [2].

For each detector, we sort the dictionary entries (clusters) into decreasing order of their MMI scores, and plot the sorted scores as a function of their position in the ordering. We call such a plot an MMI curve (see Fig.1). A detector's performance and suitability for object recognition can be measured by the Area-Under-Curve (AUC) and the shape of the MMI curve. The MMI curve for a perfect detector is a horizontal line at full score (this also gives maximum AUC). If a detector produce an MMI curve which is above average but relatively flat, such as the MMI curve for *DoG* on the cars dataset in Fig.1, it indicates that most of the detected regions are fairly distinctive and discriminative, only a few of the detections are very noisy. Under this situation, the classifiers that assume equal contribution from all features such as Nearest Neighbor and Neural Networks are probably able to tolerate the noise and give high recognition performance. On the other hand, if a detector generates a curve which has very high scores for the top ranking clusters but relatively low scores for the following clusters, for example, the MMI curve for *Curvilinear* detector on stoneflies in Fig.1. It shows that the detector can only find a few highly distinctive and discriminative regions while at the same time producing many uninformative detections. In this case, the classifiers mentioned above will probably fail. While this detector may work well with the algorithms based on discriminative feature selection [4,6], which are able to achieve high classification accuracy using only a small part of relevant features. In summary, MMI curves are valuable guidelines for the selection of detectors and the design of the image classifiers.

3. Evaluation results

We experimented with three benchmark object recognition datasets: Caltech [2,4], GRAZ [6] and Stoneflies [2]. Due to space limitation, we show the evaluation results on six object classes. Some objects are highly textured (e.g. leopards), some are structured (e.g. leaves); and these datasets differ greatly in their complexity. Each experiment is repeated 10 times with random selection of the clustering and evaluation sets, and the results are the average of 10 iterations. The MMI curves are shown in Fig.1.

We can see that all the detectors work fairly well on simple objects, such as leaves; more than half of the dictionary entries have mutual information above 0.15, and some entries achieve very high mutual information scores (> 0.5). But for relatively complex problems, such as leopards and stoneflies, the performance of detectors inevitably degrades a lot; most of the entries have the mutual information scores below 0.05.

Different detectors exhibit different characteristics and performance for different object classes, which are illustrated by the shapes of their MMI curves.

DoG has best overall performance for most of the datasets (except leaves set). This demonstrates *DoG*'s ability to detect discriminative regions in natural scenes, and its robustness to various planar transformations and limited view changes. In Fig.1, we can see that on leopards set *DoG* performs far better than other detectors. Fig.2 (a) shows the top 10 dictionary entries extracted by *DoG* on a leopard image. We can see that they are located on spots in the skin which are characteristic for leopards. On leaves set, *DoG* is outperformed by *Hessian* detectors and *Curvilinear* detector.

Curvilinear detector has evaluation scores above average on all the object classes. It works especially well on highly structured objects, such as leaves and cars. *Curvilinear* is usually able to find several highly distinctive and discriminative patterns, e.g. on leopards and stoneflies set. So implies its potential utility with feature selection methods.

On most of the datasets, *HarLap* and *HesLap* have similar MMI curves; so for *HarAff* and *HesAff*. This can be explained by their similar local intensity based detecting principles. But on leaves set in Fig.1, Hessian detectors evaluated much higher than the corresponding Harris detectors, *HesLap* works much better than all the other detectors. The top 10 dictionary entries extracted by *HesLap* on a leave image are show in Fig.2 (b). They are located on the edge of the leave which is characteristic for the object class. Few background detections are evaluated high by our method.

In addition to evaluate the relative performance of detectors, MMI curves also reveal the intrinsic characteristics of the visual dictionaries. MMI curves for *DoG* are fairly good and quite flat; it indicates that most of the dictionary entries are informative, so they can be appropriately used with the classifiers that assume equal contribution from all features; *Curvilinear* has similar MMI curve on cars set, so for Hessian detectors on leaves set. On the other hand, we also notice that some other detectors produce quite different MMI curves on some datasets. *MSER* is not stable on all the object classes in the sense that most of the entries are evaluated relatively low, while it has the ability to extract a few highly distinctive and discriminative entries in cars set. As shown in Fig. 1, its MMI curve start at 0.82, which is about 0.3 higher than any other dictionary entries. Similarly for *Curvilinear* detector and *HesAff* on Stoneflies set. Their MMI curves all start with very high score while soon drop down with the noise detections. For these dictionaries, classifiers based on discriminative feature selection are more promising. So even a detector fail to give stable detections (low repeatability), it is still possible that it can produce a small number of highly distinctive and discriminative dictionary entries if it fits the object class. In summary, the characteristics of the visual dictionaries generated by different interest region detectors can be explored directly by the shapes of their MMI curves.

We also extensively studied the robustness of our MMI evaluation criteria to several key factors: (1) density of detection; (2) the size of regions and (3) number of clusters. The MMI criterion is robust to these factors in that the relative ranking of the detectors are mostly invariant to different settings. For example, we show the evaluation results on *leaves* set with K=20 in Fig 3. Comparing with the curves in Fig 1, we can see little difference between the rankings.

To validate the MMI evaluation results, we also employ a boosted feature selection classifier [6] to select the highly evaluated dictionary entries, and test their combinational classification accuracy on real-world problem. The *evaluation* set is divided into two non-overlapping sets. One set is used as *training* set to train the image classifier; the other is used for *testing*. Then decision stumps are learned on the *training* set similarly as in Sec 2.2. Each iteration of AdaBoost searches among the unused entries and select the one which have highest MMI score. The boosted decision stumps are then applied to testing images to evaluate the performance of the detector. The results are summaries in Table.1. We can see that the results are consistent to the comparison of the detectors using MMI curves.

Table 1. Classification accuracies of detectors (%)

Class	*Har-Lap*	*Hes-Lap*	*Har-Aff*	*Hes-Aff*	*DoG*	*MS-ER*	*Cur-vi*
Lea	98.9	**99.6**	98.5	99.3	99.3	92.7	**99.6**
Car	96.4	96.4	96.4	94.1	**98.3**	97.2	93.6
Face	97.05	97.9	96.8	98.8	**99.7**	99.1	**99.7**
Leop	80.65	80.6	79.6	80.7	79.4	80.8	**82.1**
Bike	72.05	**75.6**	71.6	67.3	70.3	69.4	61.3
SF	80.8	78.7	70.2	80.9	**83.0**	70.2	69.4

4. Conclusions

In this paper, we proposed MMI curves to evaluate the discriminative power of the visual dictionaries built from different interest region detectors. Extensive experiments are performed on benchmark datasets.

References

[1] K. Mikolajczyk and C. Schmid, "Scale & Affine Invariant Interest Point Detectors," *IJCV*, 2004.

[2] W. Zhang et. al. A hierarchical object recognition system based on multi-scale principal curvature regions, *ICPR*, 2006.

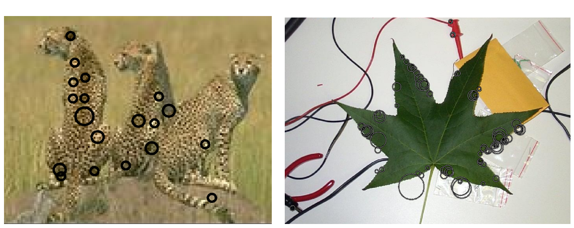
[3] D. G. Lowe. "Distinctive image features from scale-invariant keypoints". *IJCV*, 2004.

[4] G. Dorko and C. Schmid. "Object class recognition using discriminative local features". *PAMI*, submitted 2004.

[5] K. Mikolajczyk, B. Leibe and B. Schiele. Local features for object class recognition. *ICCV*, 2005.

[6] A. Opelt et. al. "Generic object recognition with boosting". *PAMI*, 2006.

[7] J. Matas et.al. Robust wide-baseline stereo from maximally stable extremal regions. In *Image and Vision Computing*, 2004.



(a) (b)

Figure 2. Examples of highly evaluated regions

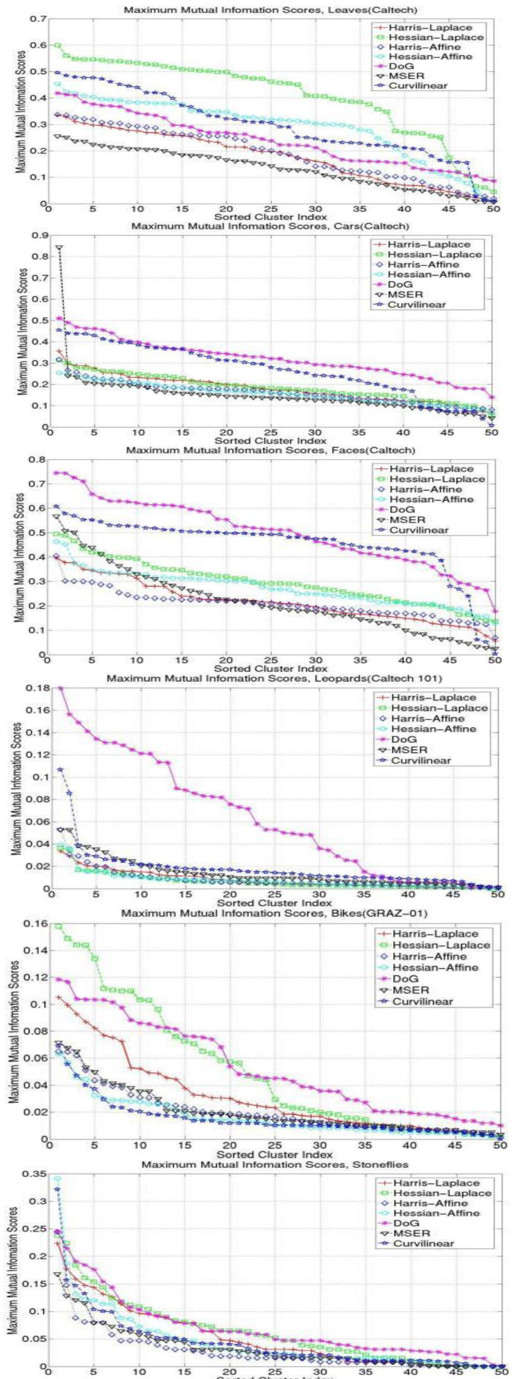


Figure 1. MMI curves of detectors

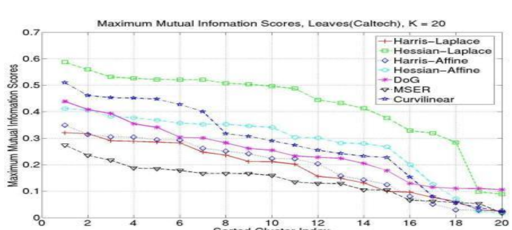


Figure 3. MMI curves of detectors with *K* = 20