Inter-cluster features for medical image classification

Siyamalan Manivannan, Ruixuan Wang, Emanuele Trucco

CVIP, School of Computing, University of Dundee, UK
{msiyamalan, ruixuanwang, manueltrucco}@computing.dundee.ac.uk

Abstract. Feature encoding plays an important role for medical image classification. Intra-cluster features such as bag of visual words have been widely used for feature encoding, which are based on the statistical information within each clusters of local features and therefore fail to capture the inter-cluster statistics, such as how the visual words co-occur in images. This paper proposes a new method to choose a subset of cluster pairs based on the idea of Latent Semantic Analysis (LSA) and proposes a new inter-cluster statistics which capture richer information than the traditional co-occurrence information. Since the cluster pairs are selected based on image patches rather than the whole images, the final representation also captures the local structures present in images. Experiments on medical datasets show that explicitly encoding inter-cluster statistics in addition to intra-cluster statistics significantly improves the classification performance, and adding the rich inter-cluster statistics performs better than the frequency based inter-cluster statistics.

1 Introduction

The Bag-of-Words (BoW) approach is widely applied as a feature encoding method for medical [1] as well as natural [2, 3] image classification. In BoW, firstly local features such as SIFT [4] extracted from training images are used to build a dictionary. This dictionary represents a set of visual words (or clusters) which are then used to compute a BoW frequency histogram as a feature vector for any give image. BoW captures the *intra-cluster* statistics of each cluster by just counting the number of local features falling into that cluster (0^{th}-order statistics). On the other hand, *VLAD* [5] and Fisher Vector (FV) [6] represents the intra-cluster information by a rich statistical representation compared to BoW. In VLAD a distance measure between the cluster center and the local features which are assigned to that cluster is used as the intra-cluster information (1^{st}-order statistics). In addition to the 0^{th} and 1^{st} order statistics, FV also considers 2^{nd} order statistics (i.e., variance for each feature component) [6] *within* each cluster. All the above encoding methods (BoW, VLAD and FV) consider that local features extracted from images are independent to each other and none of them captures (1) the *inter-cluster* statistical information (e.g., how two visual words co-occur in each image) and (2) the local structure information of images.

To capture inter-cluster information, co-occurrences between all pairs of visual words are considered as features for classification [2,3]. However, this leads to a very high-dimensional feature vector. Including inter-cluster features from pairs of clusters which do not have relevant information for classification may decrease classification performance. Recently a mutual information based criterion has been used to select cluster pairs whose co-occurrence information was then used for classification [7]. However, all these methods [2,3,7] only consider the dependency between two visual words (first-order co-occurrence) and failed to consider any higher-order dependencies (discussed in section 2). The inter-cluster information in these methods is represented merely as the number of co-occurrence between two clusters. In contrast, we make use of higher-order co-occurrence information to select the informative cluster pairs and encode the inter-cluster features using a richer representation. The contributions of this paper include:

- A new method to select a subset of cluster pairs based on Latent Semantic Analysis (LSA) by considering higher-order co-occurrence of visual words.
- A patch-based method to construct the term-document matrix in the LSA framework, which can capture structural information of objects in images.
- A new inter-cluster feature to capture rich statistical information between selected pairs of clusters, which performs better than co-occurrence frequency.
- Experimental evidence showing that adding inter-cluster statistics (even from a small subset of cluster pairs) improves medical image classification.

2 Inter-cluster features

This section focuses on adding inter-cluster statistical information to intra-cluster statistics (e.g., BoW) to represent images. A new method is proposed to choose a subset of cluster pairs by considering the higher-order co-occurrence of visual words within local image regions and introduces an inter-cluster feature which captures rich statistical information between any chosen cluster pairs.

2.1 Selection of cluster pairs based on LSA

Latent Semantic Analysis (LSA) is a well-known technique applied to a wide range of tasks such as search and retrieval [8] and classification [9]. Let $\mathbf{A}$ be a *term-document matrix* with t rows (terms) and d columns (documents), where the element $A(i,j)$ represents the frequency of the occurrence of term i in document j. In image analysis domain, terms correspond to visual words and documents often (but not always, see Section 2.2) correspond to images. In this paper terms and words are used interchangeably. An example of term-document matrix is shown in Figure 2. In LSA, a low-rank (e.g., rank-k) approximation $\mathbf{A}_k$ of matrix $\mathbf{A}$ is obtained by keeping the k largest non-zero singular values in the SVD of $\mathbf{A}$ ($\mathbf{A} = \mathbf{TSD}^{\mathrm{T}}$), i.e., $\mathbf{A}_k = \mathbf{T}_k\mathbf{S}_k\mathbf{D}_k^{\mathrm{T}}$, where the t-by-k matrix $\mathbf{T}_k$, the k-by-k diagonal matrix $\mathbf{S}_k$, and the d-by-k matrix $\mathbf{D}_k$ are respectively the truncated versions of the original matrices **T**, **S**, and **D**. Then the i-th row in $\mathbf{T}_k\mathbf{S}_k$ can be used to represent the semantic meaning of the i-th term (or word) in the so-called k-dimensional latent semantic space, where noise can be largely suppressed by discarding the smaller singular values in **S**. Based on such semantic representation of terms, the similarities (correlations) between terms can be captured by the term-term (co-occurrence) matrix, $\mathbf{C}_k = \mathbf{T}_k\mathbf{S}_k(\mathbf{T}_k\mathbf{S}_k)^{\mathsf{T}}$ [10], where each element $C_k(i, j)$ represents the similarity between the i-th and the j-th terms, with higher positive value representing stronger similarity (or positive correlation) between terms and the lower negative value representing stronger anti-similarity (or negative correlation) between terms.

More importantly, it has been shown that term-term matrix $\mathbf{C}_k$ from the truncated matrix $\mathbf{T}_k\mathbf{S}_k$ can additionally capture *higher-order co-occurrence* information (Figure 1) between terms compared to the original co-occurrence matrix (i.e. a matrix where each element (i, j) represents how many times the words i and j co-occur in a document) which is obtained directly from documents [10]. As shown in Figure 1, terms t_1 and t_2, t_2 and t_3, and t_3 and t_4 respectively co-occur in three different documents. With the original co-occurrence matrix, only the first order co-occurrence was captured and therefore the similarity between terms t_1 and t_3 (also t_2 and t_4, and t_1 and t_4) will be zero. But there is a relationship between t_1 and t_3 via t_2. Such higher-order co-occurrence can be captured by the term-term matrix $\mathbf{C}_k$ where the corresponding entries won't be zero.

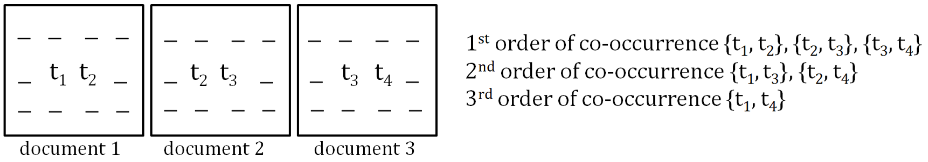

t1 t2
t2 t3
t3 t4
document 1
document 2
document 3
1st order of co-occurrence {t1, t2}, {t2, t3}, {t3, t4}
2nd order of co-occurrence {t1, t3}, {t2, t4}
3rd order of co-occurrence {t1, t4}

Fig. 1: High-order co-occurrence.

We propose to select a subset (say, P percent) of cluster (or term) pairs which have corresponding larger values in the term-term matrix $\mathbf{C}_k$. As explained above, the use of the truncated term-term matrix $\mathbf{C}_k$ instead of the original co-occurrence matrix can help choose the cluster pairs which are semantically similar. In addition, by using a small subset of cluster pairs for inter-cluster feature extraction, richer (in general with higher-dimensional) inter-cluster statistics can be extracted from the selected pairs. Instead, if all the cluster pairs are used for inter-cluster feature extraction as in [2], richer inter-cluster statistics will make feature dimensionality too high to be practically applicable for classifier training.

2.2 Construction of term-document matrix

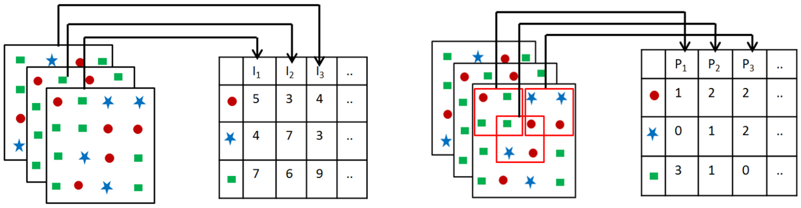
Note that the truncated term-term matrix $\mathbf{C}_k$ is obtained from the term-document matrix **A**. To construct **A**, in general, each image corresponds to one document and the occurrence of each visual word is counted within the whole image (Figure 2left). However, such term-document matrix construction does not consider any spatial relationship (e.g., far from or close to each other) between the corresponding image regions to any two visual words. As a result, the term-term matrix $\mathbf{C}_k$ won't contain any information about the spatial relationships between any two visual words. In order to make $\mathbf{C}_k$ contain certain spatial relationship between visual words, here we propose to use each image patch (with certain size) as one document (Figure 2right). In this way, the term-term matrix only considers the co-occurrence information between visual words whose corresponding image regions are within the same image patches (therefore close to each other in the image). By selecting word pairs (i, j) whose corresponding absolute values of $C_k(i, j)$ are larger in the patch-based term-term matrix $\mathbf{C}_k$, we expect that the selected highly co-occurred word pairs within image patches (i.e., local image regions) will capture certain structural information of objects in an image, e.g., teeth and nose in radio-graphic images of head often close to each other and therefore more likely appear within an image patch. The statistical information between such cluster (word) pairs may implicitly convey such structural information which cannot be captured within each cluster. What's more, the patch-based term-term matrix $\mathbf{C}_k$ can also capture the larger-scale structural information (if existing) by the higher-order co-occurrence information within $\mathbf{C}_k$, e.g., eye balls with teeth via nose.

2.3 Inter-cluster statistics

After selecting a subset of word (or cluster) pairs, we need to extract the inter-cluster information based on these pairs. Let W denote the dictionary which contains N visual words $\{\mathbf{w}_i\}$, and Π denote the selected subset of word pairs. Given any image, a number of L local descriptors (e.g., SIFT) $X = \{\mathbf{x}_l, l = 1, \ldots, L\}$ will be extracted from each image patch. Let cluster C_i denote the subset of X such that the nearest visual word for each $\mathbf{x}_l$ in C_i is $\mathbf{w}_i$. We consider the following two measures to respectively capture this inter-cluster statistics:

1. Co-occurrence of visual words: A simple measure of how many times a pair of visual words co-occur locally in each image. Consider an image patch within which visual word $\mathbf{w}_i$ occurs a times and visual word $\mathbf{w}_j$ occurs b times, and the word pair (i, j) is in the selected subset Π. The co-occurrence statistics $f(i, j)$ of these two visual words inside the image patch will be $f(i, j) = \min(a, b)$.

2. Statistical difference between two clusters: For each cluster C_i, the VLAD descriptor $\mathbf{v}_i$ is first computed as [5] $\mathbf{v}_i = \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{w}_i)$. Then for every word pair (i, j) in Π, the inter-cluster statistics is computed as $\mathbf{f}(i, j) = ||\frac{\mathbf{v}_i}{\sigma_i} - \frac{\mathbf{v}_j}{\sigma_j}||^2$, where σ_i and σ_j are the standard deviations of the clusters i and j which are computed in the dictionary learning phase by considering all the training features within those clusters. $|| \cdot ||^2$ is a component-wise squared distance measure, and therefore $\mathbf{f}(i, j)$ is a vector and will contain richer statistical information than the scalar co-occurrence value.



	I_1	I_2	I_3	..
●	5	3	4	..
★	4	7	3	..
■	7	6	9	..

	P_1	P_2	P_3	..
●	1	2	2	..
★	0	1	2	..
■	3	1	0	..

Fig. 2: Term-document matrix obtained from images (left) and patches (right).

2.4 Feature encoding

Given an image, we encode the image based on both intra-cluster and inter-cluster statistics. First we compute the intra-cluster statistics using the existing approaches such as BoW or VLAD. Then we compute the inter-cluster statistics for image patches in the image as described above. Finally we apply sum pooling over all image patches for the inter-cluster statistics to obtain a feature vector which represents the inter-cluster statistical information for the whole image. The feature vector obtained based on the intra and inter-cluster statistics are normalized individually (we use the power and $L2$ normalizations as in [11]) and concatenated together as the final image descriptor.

3 Experiments

Two medical datasets were used to evaluate the proposed method for cluster pair selection and inter-cluster features. The ICPR HEp-2 cell classification dataset (ICPR[1]) contains $13,596$ gray-scale cell images from 6 classes (homogeneous, speckled, nucleolar, centromere, golgi, and nuclear membrane), with average image size about 70×70 pixels. The Image Retrieval in Medical Applications dataset (IRMA[2]) contains 15,363 anonymous radiographs from 57 classes (of various human body parts), with images resized to be no larger than 300×300. Since the number of images is very unbalanced across IRMA classes, only 20 classes were selected, each of which contains 200 images. We used one-vs-rest multi-class SVM with linear and intersection kernels [12] for classification. SVM parameters were learned using 5-fold cross-validation on the training set. The value of k is chosen such that the $\mathbf{A}_k$ keeps 95% of its column-wise variance. BoW and VLAD features are respectively used as two intra-cluster features based on the local descriptor SIFT, where for each image, dense SIFT descriptors were extracted from each small regions of size 16×16 pixels over a grid with spacing of 4 pixels along both directions, and every 7×7 neighboring regions compose one image patch (i.e., 49 SIFT features in each patch). For ICPR dataset, we applied two-fold cross-validation and report the mean per-class accuracies (MAC) over 5 runs. For the IRMA dataset 30 images per class are selected for training and the rest are used for testing; the averaged MAC over 10 iterations are reported.

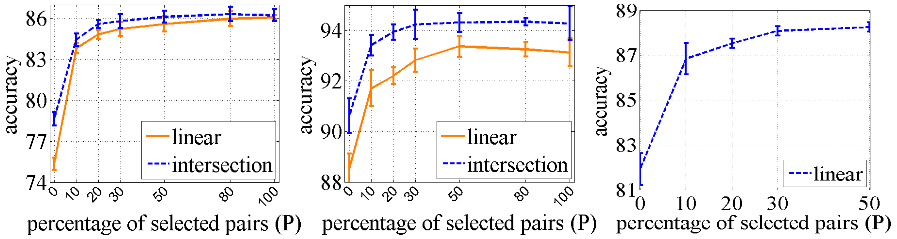
[1] http://i3a2014.unisa.it/

[2] http://ganymed.imib.rwth-aachen.de/irma/index_en.php

3.1 Effect of the inter-cluster features

When using BoW as intra-cluster feature and co-occurrence frequency of visual words as inter-cluster features, Figures 3(a)(b) show that adding inter-cluster features significantly increase the classification performance for both datasets (e.g., around 78% when $P = 0$ vs. 86% when $P > 0$ for ICPR dataset, and around 91% vs. 94% for IRMA dataset, both with dictionary size 200 and using intersection kernel). It also shows that the classification accuracy is not significantly different between selecting 10% (when $P = 10$) and all (when $P = 100$) cluster pairs, which indicates that only a small subset of cluster pairs are sufficient enough to capture the inter-cluster information. Figure 3(a)(b) also show that intersection kernel for intra-cluster feature cannot capture high-order information encoded in inter-cluster features, otherwise adding inter-cluster feature would not improve the accuracy.

Similar findings have been confirmed when using VLAD as the intra-cluster feature and the VLAD-based inter-cluster statistics for the inter-cluster features (Figure 3(c)). By comparing the classification performance from Figures 3(a) and (c), it becomes clear that, even using a smaller dictionary ($N = 32$) and a smaller subset of cluster pairs ($P = 10$ percent), VLAD plus VLAD-based inter-cluster features outperforms the corresponding BoW plus co-occurrences based inter-cluster features, i.e., 86.8% vs. 84.4% for ICPR dataset. Similar finding were found for IRMA dataset (not shown due to limited space). This indicates that both VLAD intra-cluster feature and the VLAD-based inter-cluster feature captures richer statistical information than the BoW intra-cluster feature and the co-occurrence based inter-cluster feature.



(a) ICPR dataset (b) IRMA dataset (c) ICPR dataset

Fig. 3: Effect of the inter-cluster features. $P = 0$ corresponds to intra-cluster feature, and $P > 0$ corresponds to inter-cluster feature plus intra-cluster feature. (a-b) BoW with co-occurrence, (c) VLAD with statistical cluster difference.

To further confirm the effect of inter-cluster features, in Figure 4left the sizes of the dictionaries are varied and only 20% cluster pairs are chosen based on corresponding dictionaries. It shows a significant performance improvement when adding inter-cluster features, no matter what the dictionary size is. Since

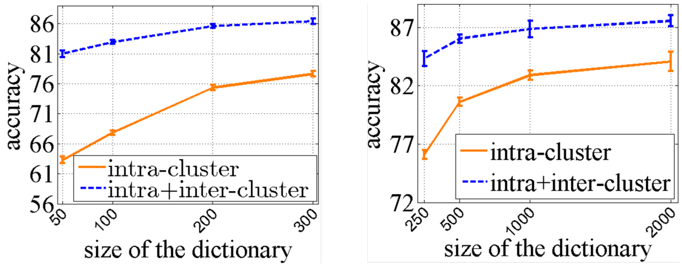


Fig. 4: Classification performance on ICPR dataset with BoW and co-occurrence based inter-cluster features using intersection kernel. See text for more details.

adding inter-cluster features for larger dictionaries tremendously increases the dimensionality of the final image representation, in another test, we capture inter-cluster features by considering only 20% pairs from a fixed small dictionary of size 100. Adding these fixed inter-cluster features to the traditional intra-cluster BOW features computed from any larger dictionary still increases the overall performance (Figure 4right). Notice that adding inter-cluster features from a fixed smaller dictionary not only increases the classification accuracy but also reduces the feature dimensionality.

3.2 Patch-based vs. image-based methods

This test is to compare the performance of patch-based with the image-based cluster pair selection for inter-cluster feature encoding on the IRMA dataset. For both methods, BoW was used as intra-cluster feature and co-occurrence of selected visual words as inter-cluster feature. The dictionary size was fixed to 200 and only 10% of pairs are selected to encode inter-cluster features. As expected, patch-based method gives the accuracy of 93.4%, much better than the accuracy 87.0% from image-based method (with standard deviation about 0.7%), supporting that patch-based method helps capture local structural information encoded in inter-cluster features.

3.3 LSA-based pair selection

In this section the LSA-based truncated term-term matrix is compared with the original co-occurrence matrix for pair selection. In this experiment a dataset containing radiographs of heads taken from four different angles collected from the IRMA dataset is considered. This dataset contains 50 images in each of the four classes. By keeping all the other factors (e.g., patch-based term-document construction and VLAD based inter-cluster feature encoding) unchanged, we found that when selecting a small subset ($P = 5$) of pairs for inter-cluster features, the pair selection based on the truncated term-term matrix performs significantly better than based on the original co-occurrence matrix (78.3% vs. 87.2%). This confirms the potential function of LSA-based pair selection in reducing noise and capturing high-order co-occurrence statistics.

3.4 Inter-cluster features for Fisher Vector

Some initial experiments with FV was also performed on ICPR dataset to observe the effect of inter-cluster features for FV. Given an image, Fisher vector $\mathbf{F_i}$ for each cluster C_i was computed based on soft-assignments (see [6] for details). The inter-cluster feature between any chosen cluster pair (i, j) was computed as $||\mathbf{F_i} - \mathbf{F_j}||^2$ (component-wise, as for VLAD). With totally 16 clusters being used, accuracy of 85.2% was obtained by FV. In comparison, adding inter-cluster features ($P = 20$) to FV significantly improves the performance to 88.7%.

4 Conclusions

This paper showed that adding inter-cluster features to the intra-cluster features significantly improves medical image classification. A new method was proposed to select a subset of cluster pairs to get the inter-cluster features. Experiments showed that adding rich inter-cluster statistics performs better than only considering the co-occurrence frequency information as the inter-cluster statistical feature. In feature work we plan to select cluster pairs based on discriminative information (i.e., class labels) and add spatial information to final representation.

Acknowledgement: This work is funded by 2011-2016 EU FP7 ERC project "CODIR: colonic disease investigation by robotic hydrocolonoscopy", collaborative between the Universities of Dundee (PI Prof Sir A Cuschieri) and Leeds (PI Prof A Neville).

References

1. Manivannan, S., Wang, R., Trucco, E., Hood, A.: Automatic normal-abnormal video frame classification for colonoscopy. In: ISBI. (2013)
2. Yang, Y., Newsam, S.: Spatial pyramid co-occurrence for image classification. In: ICCV. (2011) 1465–1472
3. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: AGIS. (2010) 270–279
4. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV. (1999)
5. Jégou, H., Douze, M., Schmid, C., Prez, P.: Aggregating local descriptors into a compact image representation. In: CVPR. (2010) 3304–3311
6. Perronnin, F., Dance, C.R.: Fisher kernels on visual vocabularies for image categorization. In: CVPR. (2007) 1–8
7. Chen, T., Yap, K.H., Chau, L.P.: From universal bag-of-words to adaptive bag-of-phrases for mobile scene recognition. In: ICIP. (2011) 825–828
8. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. JASIS **41**(6) (1990) 391–407
9. Zelikovitz, S., Hirsh, H.: Using LSI for text classification in the presence of background text. In: ICIKM. (2001) 113–118
10. Kontostathis, A., Pottenger, W.M.: A framework for understanding latent semantic indexing performance. IPM (2006) 56–73
11. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV. (2010) 143–156
12. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. JMLR (2008)