IPL at CLEF 2016 Medical Task

Leonidas Valavanis, Spyridon Stathopoulos, and Theodore Kalamboukis

Information Processing Laboratory,
Department of Informatics,
Athens University of Economics and Business,
76 Patission Str, 10434, Athens, Greece
valavanisleonidas@gmail.com,tzk@aueb.gr
http://ipl.cs.aueb.gr/index_eng.html

Abstract. In this paper we present the image classification techniques performed by the IPL Group for the subfigure classification subtask of ImageCLEF 2016 Medical Task. For the visual representation of images, various state-of-the-art visual features, such as, Bag of Visual Words computed with pyramid-histogram of-visual-words descriptors and quad-tree bag-of-colors were adopted. We present the results of our runs and our extensive experiments applying early or late fusion on the results obtained from a multi-class linear kernel support vector machine. Our top run was ranked 3rd among 34 runs.

Key words: pyramid-histogram of visual words, bag of visual words, bag of colors, early-fusion, late-fusion, textual-classification, support vector machines

1 Introduction

Image classification is perhaps the most important and challenging task within the field of computer vision with applications in several domains. A broad area of image-processing approaches is directed by image classification, the automated assignment of unknown images into a set of predefined categories.

In the medical domain Content Based Image Retrieval (CBIR) plays an important role in supporting diagnosis, treatment and teaching [1]. Visual image classification into a relatively small number of classes, has shown to deliver good results in several benchmarks. Approaches combining both visual and textual techniques for classification have shown to be promising in medical image classification tasks. Here we should mention the substantial contribution of the ImageCLEFmed task [2] focusing on medical images over a decade on the CBIR and classification tasks.

The ImageCLEF 2016 Medical Task, [3], consists of 5 subtasks: compound figure detection, figure separation, multi-label classification, subfigure classification, caption prediction. Subfigures extracted from compound images are classified into 30 heterogeneous classes ranging from diagnosis images to various biomedical illustrations. Some image categories were represented by few training examples, thus the enrichment of the original collection was necessary in order to counteract the imbalanced dataset. Over the past years of the contest there was a large class of compound images that contained sub-images of several modalities something which made it difficult to train a classifier. This year there are no compound images in the subfigure classification subtask. However, both, the train and the test sets remain unbalanced with one very large category (GFIG, 2085) and some other categories that contain just few images(GPLI 2) or (DSEE, 3).

This year our group participated only in the subfigure classification subtask. Details of this task can be found in the overview paper [3] and the web page of the contest [1]. Our approach to classification is based on merging two well known models, that of the BoW, [4] and a generalized version of bag of colors (BoC), [5] approach combined with early or late fusion which gave us the third best performing position.

In the next section we present a detailed description of the modelling techniques and data fusion used. In section 4, the classification tools and parameters are described as well as the submission runs with our results. Finally, Section 5 concludes our work.

2 Image Visual Representation

Inspired from text retrieval, the Bag-of-visual Words (BoW) approach has shown promising results in the field of image retrieval and classification. In this vein, we based our approach to the BoW model for the image classification task. In this section, we describe the methodology used for the visual and textual representation of images.

2.1 Pyramid Histogram of Visual Words (PHOW)

PHOW is an extension of the BoW model used for image classification. In this model, we identify small regions (local interest points) known as, salient image patches that contain rich local information of the image. To extract such key-points, the SIFT [6] or the Dense SIFT [7] descriptors are employed. However, the number of features extracted from local interest points may vary, depending on the image. In order to have a fixed number of feature dimensions, a visual codebook is created by clustering the extracted local interest points of a number of sample images, using the k-means clustering algorithm. Each cluster (visual word) represents a different local pattern, which shares similar interest points. The histogram of an image, is created by performing a vector quantization which assigns each key-point to its closest cluster (visual word) [8]. However, as it is known, the BoW model loses the spatial information of the local descriptors due to the clustering which, limits severely their discriminative power. Pyramid Histogram of Visual Words (PHOW) addresses this problem by dividing the image into increasingly fine sub-regions of equal size, which are called pyramids.

[1] `http://www.imageclef.org/2016/medical`

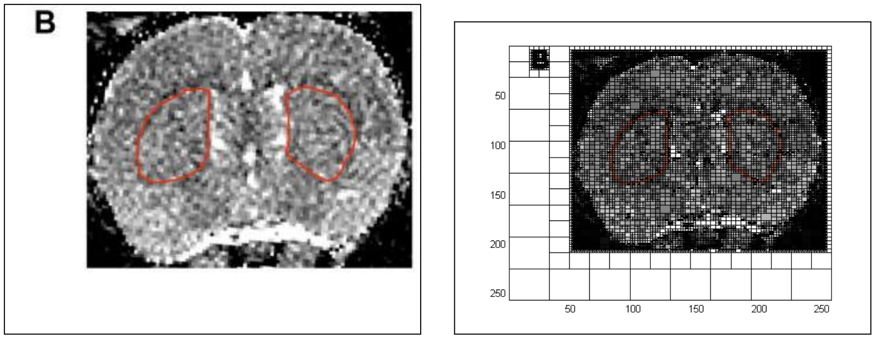
The histogram of visual words is computed in each local sub-region of the image and in the sequel they are concatenated into a single feature vector [9]. For our experiments, we partition the image into 2×2 and 4×4 sub-regions and then combine the generated quantizations. As for the size of the visual codebook, after experimentation with several values, we selected 1536 visual words. Thus each image was represented with a vector of 30720 features (2x2x1536 + 4x4x1536).

2.2 Quad-Tree Bag-of-Colors Model(QBoC)

With the BoC model [5] a color vocabulary is learned from a sub-set of the image collection. This vocabulary is used to extract the color histograms for each image. Through experiments, it has been shown that using a learned color vocabulary improves retrieval performance over a flat color space quantization. Furthermore, this model is succesfully fused with the SIFT descriptor into a compact binary signature [10] increasing further the performance of classification. The BoC model was used for classification of biomedical images in [11] and it was shown that it is combined successfully with the BoW-SIFT model in a late fusion manner. Similarly to the BoW model the main drawback with the BoC is the lack of spatial information. Furthermore, it is evident that the construction of the vocabulary and in particular the selection of its size is another weak point of the algorithm. To address this problem, we have extended the BoC model applying a quad-tree-decomposition of images [12]. Quad-Tree decomposition sub-divides an image into regions of homogeneous colors. Each time the image is split into four equal size squares and the process continues until we reach a sub-region of size 1×1 pixel (see figure 1b). To speed up the pre-processing of the images the Quad-Tree decomposition may end when we reach a sub-region of 2×2 pixels. Similar colors within a sub-region are quantized into the same color. This is tuned with an extra parameter which, was set to 0.15 in all our runs. In both models the TFIDF weights of visual words were calculated and the image vectors were normalized with the L_1 norm.

3 Textual Representation

The text representation for the sub-figure images is derived from the caption of their corresponding compound figures. The caption of a compound figure is assigned to all its constituent subfigures. This makes it difficult to distinguish between sub-images, and is a point to be improved in the future. For text retrieval we used the vector space model with TFIDF weights of the terms. While we didn't submit runs using textual information due to a misunderstanding, experimentation outside competition showned that stemming significantly drops the performance of categorization (see section 4.4).



(a) (b)

Fig. 1. Representation of image 1471-2202-11-1-4-2 (a) original image; (b) QBoC image.

4 Image Classification

4.1 Experiments Settings

All our experiments were conducted using several combinations of the two models described in section 2. For the classification the LibLinear classifier [2] was employed, an open source library for large scale linear classification [13]. Linear SVMs are in general much faster to train and predict than the non-linear and can approximate large scale non-linear SVMs using a suitable feature map. Efficient feature mapping can be achieved using additive kernels, commonly employed in computer vision, with the homogeneous kernel map being the most common [14]. The homogeneous kernel map includes the intersection, Hellinger's, Jensen Shannon, Chi2, which allows large scale training of non-linear SVMs. The transformation of the data results into a compact linear representation which reproduces the desired kernel to a very good level of approximation. This transformation makes the use of linear SVM solvers feasible [3], [4]. In our experiments, the homogeneous kernel mapping of VLFeat is used and more specifically Chi2 kernel. The implementation of VLFeat does not require any parameters but experiments have shown that results can be improved slightly by changing the Gamma parameter. The Gamma parameter sets the homogeneity degree of the kernel. The SVM model was tuned using n-fold cross validation to find the best cost. LibLinear has an embedded grid search which conducts n-fold cross validation with different costs and finds the best one. Besides from the cost parameter, that is discovered using grid search, bias multiplier and kernel type were given. Results were not greatly affected when varying bias multiplier or kernel type. After experimentation using several parameters, results yielded better performance with

[2] `https://www.csie.ntu.edu.tw/~cjlin/liblinear/`

[3] `http://www.robots.ox.ac.uk/~vgg/software/homkermap/#r1`

[4] `http://vision.princeton.edu/pvt/SiftFu/SiftFu/SIFTransac/vlfeat/doc/api/homkermap.html`

cost 10, Gamma 0.5 and the L2-Regularized L2-loss support vector classification kernel.

4.2 Early and Late Data Fusion

In early fusion also referred to as feature fusion, [15], image representation features extracted from different models are integrated into a single unified representation. Normalization techniques may be applied before the integration so that features are on the same scale. There is only one learning phase that handles all multimodal features together. Five of our submitted runs were conducted using early fusion. In late fusion also referred to as decision level fusion, multiple probabilistic output scores obtained from separate classifiers are combined into a single vector to form the final decision. Models are trained and classified separately and their respective outputs are combined to form the final decision. In contrast to the early fusion, late fusion requires two learning phases and there is a potential loss of correlation in the mixed feature space. Nevertheless, late fusion does not suffer from the integration problem early fusion does and can be easily used due to its simplicity and scalability. Five of our submitted runs were conducted using late fusion.

4.3 Submitted Runs and Results

In this year's contest we submitted ten visual runs for the subfigure classification subtask. The results are presented in table 1. Early tests on the learning curves of our model on the imageCLEF 2013 dataset shown that the test error drops continuously with increasing the training instances. This suggests that with a larger dataset the test error would drop even more. Thus we have enriched the poorest train categories with new images. These categories were the following 14/30: DRAN, DRCO, DRCT, DRPE, DRUS, DRXR, DSEC, DSEE, DSEM, DVDM, DVEN, GFLO, GMAT, GPLI. Thus we have used two datasets with our runs:

- Original Dataset: The original training collection distributed for the subgure classication task in ImageCLEF2016 Medical task containing 6776 images and the
- Enriched Dataset: The original training collection was enriched with 482 images from the ImageCLEF 2013 Modality Classication training collection [16]. The enriched dataset contains 7258 images .

The name of each run describes the methods and the parameters used in the run. For example, the first run in table 1, corresponds to an early fusion experiment on the enriched dataset of the

- QBoC model, using a quad tree decomposition of the image terminating at a block of size 1 × 1 and a codebook of 256 colors in the RGB color space and the

– BoW model, with the default PHOW 2 -level descriptor with 1536 features. A color option is used to compute the color variant of the descriptor, i.e. RGB. The value of the parameter ”default”, denotes that the gray-scale variant of the descriptor is computed.

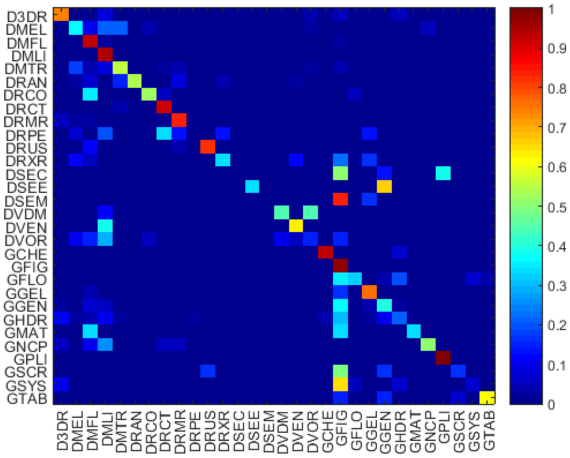


Fig. 2. Confusion matrix.

4.4 Results

From the confusion matrix, in figure 2, corresponding to our first run, we see that in three categories, zero true positive examples were assigned. These categories were: the PET (DRPE) where the majority of the examples were classified into Computerized Tomography (DRCT) and the Electrocardiogpaphy (DSEC) and Electromyography (DSEM) categories where most of the examples were classified as statistical figures, graphs and charts (GFIG). These three categories happen to have the smallest learning sets even after the enrichment with new images with 30, 39, and 23 train images each.

Although we submitted runs exclusively for visual categorization, for completeness, we present here our results for textual and mixed classification. Our textual representation of images was based on a naive TFIDF bag of words model with stopword removal and stemming. The textual classification on the enriched dataset attained an accuracy of 63.68% with stemming and 70.07% without stemming. Our mixed results combining QBoC with PHOW and Text in an early fashion mode, with weights $(0.5, 0.3, 0.2)$ respectively, attained accuracy 86.9%.

Run_ID	**Accuracy**
SC_enriched_GBOC_1x1_256_RGB_Phow_Default_1500_EarlyFusion	84.01
SC_enriched_GBOC_1x1_128_HSV_Phow_RGB_1500_EarlyFusion	83.46
SC_enriched_GBOC_1x1_128_HSV_Phow_RGB_1500_LateFusion	82.66
SC_original_GBOC_1x1_256_RGB_w_0.6_Phow_Default_1500_w_0.4_EarlyFusion	81.73
SC_original_GBOC_1x1_256_RGB_Phow_Default_1500_EarlyFusion	81.70
SC_original_GBOC_1x1_128_RGB_Phow_Default_1500_EarlyFusion	81.32
SC_original_GBOC_1x1_256_RGB_Phow_Default_1500_LateFusion	80.17
SC_original_GBOC_1x1_128_HSV_Phow_RGB_1500_LateFusion	80.14
SC_original_GBOC_1x1_128_RGB_Phow_Default_1500_LateFusion	79.45

Table 1. IPL submitted visual runs on subfigure classification.

5 Conclusions

In this paper we presented the image classification techniques performed by the IPL Group for the subfigure classification subtask at ImageCLEF 2016 Medical Task. For our runs, we used Early and Late Fusion on two bag-of-visual-words models. The first model was a novel generalized version of the BoC model, and the second was the classical BoW with the PHOW descriptor to represent images. Our experiments show that using Early or Late Fusion performs better than any of the two models on their own. Providing visual image representation with textual representation, proved to be beneficial for classification accuracy. The results so far with our new approach of the QBoC model are encouraging and several new directions have emerged which need further investigation.

References

1. Müller, H., Michoux, N., Bandon, D., Geissbühler, A.: A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. I. J. Medical Informatics **73**(1) (2004) 1–23
2. Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems– an overview of the medical image retrieval task at ImageCLEF review–2014. Computerized Medical Imaging and Graphics (2014)
3. García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 medical task. In: Working Notes of CLEF 2016 (Cross Language Evaluation Forum). (September 2016)
4. Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR (2). (2005) 524–531
5. Wengert, C., Douze, M., Jégou, H.: Bag-of-colors for improved image search. In: Proceedings of the 19th International Conference on Multimedia 2011, Scottsdale, AZ, USA, November 28 - December 1, 2011. (2011) 1437–1440
6. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2. ICCV '99, Washington, DC, USA, IEEE Computer Society (1999) 1150–1157
7. Bosch, A., Zisserman, A., Muñoz, X.: Image classification using random forests and ferns. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007. (2007) 1–8

8. Yang, J., Jiang, Y.G., Hauptmann, A.G., Ngo, C.W.: Evaluating bag-of-visual-words representations in scene classification. In: Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval. MIR '07, New York, NY, USA, ACM (2007) 197–206
9. Khaligh-Razavi, S.: What you need to know about the state-of-the-art computational models of object-vision: A tour through the models. CoRR **abs/1407.2776** (2014)
10. Jégou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. International Journal of Computer Vision **87**(3) (2010) 316–336
11. de Herrera, A.G.S., Markonis, D., Müller, H.: Bag–of–colors for biomedical document image classification. In: Medical Content-Based Retrieval for Clinical Decision Support. Springer (2013) 110–121
12. Yin, X., Düntsch, I., Gediga, G.: Quadtree representation and compression of spatial data. Springer (2011)
13. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. J. Mach. Learn. Res. **9** (June 2008) 1871–1874
14. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. IEEE Trans. Pattern Anal. Mach. Intell. **34**(3) (March 2012) 480–492
15. Zhou, X., Depeursinge, A., Müller, H.: Information fusion for combining visual and textual image retrieval in imageclef@icpr. In: Proceedings of the 20th International Conference on Recognizing Patterns in Signals, Speech, Images, and Videos. ICPR'10, Berlin, Heidelberg, Springer-Verlag (2010) 129–137
16. De Herrera, A., Kalpathy-Cramer, J., Fushman, D., Antani, S., Müller, H. In: Overview of the ImageCLEF 2013 medical tasks. Volume 1179. CEUR-WS (2013)