ImageCLEF 2010 Modality Classification in Medical Image Retrieval: Multiple feature fusion with normalized kernel function

Xian-Hua Han[1], Yen-Wei Chen[1]

College of Information Science and Engineering, Ritsumeikan University, Kasatsu-shi, 525-8577,Japan.

Abstract. In this paper, we describe an approach for the automatic modality classification in medical image retrieval task of the 2010 CLEF cross-language image retrieval campaign (ImageCLEF). This work is focused on the process of feature extraction from medical images and fusion the different extracted visual feature and textual feature for modality classification. To extract visual features from the images, we used histogram descriptor of edge, gray or color intensity and block-based variation as global features and SIFT histogram as local feature, and the binary histogram of some predefined vocabulary words for image captions is used for textual feature. Then we combine the different features using normalized kernel functions for SVM classification. The proposed algorithm is evaluated by the provided modality dataset by ImageCLEF2010.

1 Introduction

Imaging modality is an important aspect of the image for medical retrieval [1-6]. In user-studies, clinicians have indicated that modality is one of the most important filters that they would like to be able to limit their search by. Many image retrieval websites (Goldminer, Yottalook) allow users to limit the search results to a particular modality. However, this modality is typically extracted from the caption and is often not correct or present. Studies have shown that the modality can be extracted from the image itself using visual features [7,8, 9]. Therefore, In this paper, we propose to use both visual and textual features for medical image representation, and combine the different features using normalized kernel function in SVM.

In computer vision, studies have shown that the simple global features such as histogram of edge, gray or color intensity and so on can represent images, and give the acceptable performance in image retrieval or recognition research fields. Based on the success of the above mentioned visual features for general image recognition, we also use them as medical image representation for modality classification. Recently, using local visual feature for image representation has been become very popular, and been proved to be very effective for image categorization or retrieval [10]. The most famous approach for image representation using local visual feature is bag of keypoints [11, 12]. The basic idea of bag of keypoints is that a set of local image patches is sampled using some method (e.g. densely, randomly, or using a keypoint detector) and a vector of visual descriptors is evaluated on each patch independently (e.g. SIFT descriptor,

normalized pixel values). The resulting distribution of descriptors in descriptor space is then quantified in some way (e.g. by using vector quantization against a pre-specified codebook to convert it to a histogram of votes for (i.e. patches assigned to) codebook centres) and the resulting global descriptor vector is used as a characterization of the image (e.g. as feature vector on which to learn an image classification rule based on an SVM classifier). Furthermore, according to the visual properties of medical images, we also calculate a histogram of small-block variance as visual feature for image representation. For textual feature, we pre-define 90 vocabulary words somewhat according to the statistical properties of training samples' captions and our (not radiologist) knowledge about medical modality, and calculate a binary histogram for any medical image using their captions. After obtain the different feature for image representation, we combine them together using kernel function for SVM classifier. Because different features maybe have deferent scale and dimension, in order to allow each individual feature to contribute equally for modality classification, we normalize the distance between two samples using mean distance of all training samples, and then, obtain the kernel function for each individual feature. The final kernel for SVM classification is the mean of individual kernel, which can be called Joint Kernel Equal Contribution (JKEC). The proposed algorithm is evaluated on the modality dataset of ImageCLEF2010, and the classification rate is almost approximated the classification goal of the modality classification task.

2 Feature extraction for image representation

In this section we describe how we extract a feature representation which is somewhat robust to the high variability inherent in medical images and includes enough discriminative information for modality category. As we known that it is difficult to classify image categorization only with one type of image feature. So in this paper, we represent images with differen images features: including gray and color intensity histogram, block-based edge and variance histogram and popular bag-of-words model as visual feature, and a binary histogram of the predefined vocabulary words for image captions as textual feature. Then we merge them together for modality classification. Next, we simply introduce the used features for medical image representation.

2.1 Visual features

Gray and Color intensity histogram: Intensity histograms are widely used to capture the distribution information in an image. They are easy to compute and tend to be robust against small changes of camera viewpoints. For Gray intensity histogram, we can calculate the number of each intensity (0–255) for all image pixel, and normalize it using pixel number. Given an image $\mathbf{I}$ in some color space (e.g., red, green, blue), for calculate color histogram the color channels are quantized into a coarser space with k bins for red, m bins for green and l bins for blue. Therefore the color histogram is a vector $\mathbf{h} = (h_1, h_2, \cdots, h_n)^T$, where $n = kml$, and each element h_i represents the number of pixels of the discretized color in the image. We assume that all images have been scaled to the same size. Otherwise, we normalize histogram elements as

$$h_j^{'} = \frac{y_j}{\sum_{j=0}^{n} y_j} \tag{1}$$

$\mathbf{Block-based\ edge\ histogram}$: We firstly segment the image into several blocks, and calculate edge histogram weighted by gradient intensity in each block. In experiment, we grid-segment an image into 4 by 4 block, and calculate a 20-bin edge histogram in each block. So we have 320 (20*16)-dimensional edge histogram feature for medical image representation.

$\mathbf{Block-based\ variace\ histogram}$: For each pixel in an image, a small patch centered by the specific pixel are used for calculating the local variation of the pixel. after obtaining the local variation of all pixels in the image, a histogram of variation intensity is calculated for the image representation.

$\mathbf{bag-of-words\ feature}$: In computer vision, local descriptors (i.e. features computed over limited spatial support) have proved well-adapted to matching and recognition tasks, as they are robust to partial visibility and clutter. In this paper, we use grid-sampling patches, and then compute appearance-based descriptors on the patches. In contrast to the interest points from the detector, these points can also fall onto very homogeneous areas of the image. After the patches are extracted,the SIFT [10] descriptor is applied to represent the local features. The SIFT descriptor computes a gradient orientation histogram within the support region. For each of 8 orientation planes, the gradient image is sampled over a 4 by 4 grid of locations, thus resulting in a 128-dimensional feature vector for each region. A Gaussian window function is used to assign a weight to the magnitude of each sample point. This makes the descriptor less sensitive to small changes in the position of the support region and puts more emphasis on the gradients that are near the center of the region. To obtain robustness to illumination changes, the descriptors are made invariant to illumination transformations of the form $aI(x) + b$ by scaling the norm of each descriptor to unity [10]. These SIFT features are then clustered with a k-means algorithm using the Euclidean distance. Then we discard all information for each patch except its corresponding closest cluster center identifier. For the test data, this identifier is determined by evaluating the Euclidean distance to all cluster centers for each patch. Thus, the clustering assigns a cluster $c(x)21, ...C$ to each image patch x and allows us to create histograms of cluster frequencies by counting how many of the extracted patches belong to each of the clusters. The histogram representation $h(X)$ with C bins is then determined by counting and normalization such that:

$$h_c(X) = \frac{1}{L_X} \sum_{l=1}^{L_X} \delta(c, c(x_l)) \tag{2}$$

where δ denotes the Kronecker delta function. Figure 2 shows the procedure bag-of-words(BoW) feature extraction and the extracted histogram feature of example images. Obviously, there exist alternatives to algorithmic choices made in the proposed method. For example, different interest point detectors can be used. However, it do not manifest obvious merit for different background cluster of images. Furthermore, the geometrical relation between the extracted patches is completely neglected in the approach presented here. While this relation could be used to improve classification accuracy, it

remains difficult to achieve an effective reduction of the error rate in various situations by doing so.

2.2 Textual features

According to the statistical properties of word occurrence in each training modality image's captions and our prior knowledge about the classifying modality, we select 90 key-words, such as CT, Curve, MR, urethrogram, PET and so on, as the vocabulary for forming a binary histogram for each medical image. The binary histogram for image representation is 90-dimension vector, where each dimension is correspond to one selected keyword. If one key-word is appeared one or more than one times in an image's caption, the value of the corresponding dimension in its represented binary histogram will be 1, otherwise it will be 0.

3 feature fusion

Given a training set $(x_i, y_i)_{i=1,2,\cdots,N}$ of N instances consisting of an image $x_i \in \chi$ and a class label $y_i \in 1, 2, \cdots, C$, and given a set of F image features f_m: $\chi \rightarrow \Re^{d_m}$, $m = 1, 2, \cdots, F$, where d_m denotes the dimensionality of the m^{th} feature, the problem of learning a classification function $y : \chi \rightarrow 1, 2, \cdots, C$ from the features and training set is called feature combination problem. In computer vision, the problem of learning a multi-class classifier from training data is often addressed by means of kernel methods. Kernel methods make use of kernel functions defining a measure of similarity between pairs of instances. In the context of feature combination it is useful to associate a kernel to each image feature as the following Eq. 3, and combine the kernels of different features together. For a kernel function K of each feature between real vectors we define the short-hand notation:

$$K_m(\mathbf{I}_i, \mathbf{I}_j,) = K(f_m(\mathbf{I}_i), f_m(\mathbf{I}_j)) = K(S(f_m(\mathbf{I}_i), f_m(\mathbf{I}_j))) \tag{3}$$

where $\mathbf{I}_i$ and $\mathbf{I}_j$ are two samples, $f_m(\mathbf{I}_i)$ is the m^{th} extracted feature from the sample $\mathbf{I}_i$ and $S(f_m(\mathbf{I}_i), f_m(\mathbf{I}_j))$ is the similarity measure between the m^{th} features of the samples $\mathbf{I}_i$ and $\mathbf{I}_j$. Then the image kernel K_m: $\chi \times \chi \in \Re$ only considers similarity with respect to image feature f_m. If the image feature is specific to a certain aspect, say, it only considers color information, then the kernel measures similarity only with regard to this aspect. The subscript m of the kernel can then be understood as indexing into the set of features. Because different features maybe have deferent scale and dimension, in order to allow each individual feature to contribute equally for modality classification, we normalize the distance between two samples using mean distance of all training samples, and then, obtain the kernel function for each individual feature f_m. The final kernel for SVM classification is the mean of individual kernel, which can be called Joint Kernel Equal Contribution (JKEC). For the feature similarity calculation of two simples, we use χ^2 distance as the following:

$$S(f_m(\mathbf{I}_i), f_m(\mathbf{I}_j)) = \sum_{1}^{L} \frac{(x_l - y_l)^2}{x_l + y_l} \tag{4}$$

where $\mathbf{x}$ and $\mathbf{y}$ represent the m^{th} features $f_m(\mathbf{I}_i), f_m(\mathbf{I}_j)$ of samples i and j, respectively, and x_l is the l^{th} element of the vector $\mathbf{x}$. Then, the RBF function is used for calculating the kernel:

$$K_m(\mathbf{I}_i, \mathbf{I}_j,) = \exp(\frac{-S(f_m(\mathbf{I}_i), f_m(\mathbf{I}_j))}{\gamma}) \tag{5}$$

where γ is the normalized item for Joint Equal Contribution of each feature. Here, we use the distance mean of all training samples as γ, which will lead to similar contribution of each feature to kernel. The proposed algorithm is evaluated on the modality dataset of ImageCLEF2010, and the classification rate is almost approximated the classification goal of the modality classification task.

4 Experimental setup

4.1 Image Data

The database released for the ImageCLEF-2010 Medical modality classification in medical retrieval task includes 2390 annotated modality images (CT: 314; GX: 355; MR: 299; NM: 204; PET: 285; PX: 330; US: 307; XR:296) for training and a separate evaluated set consisting of 2620 images. The aim is to automatically classify the evaluated set using 8 different modality label sets including CT, MR, PET and so on. some example images are shown in Fig. 1. A more detailed explanation of the database and the tasks can be found in [13].

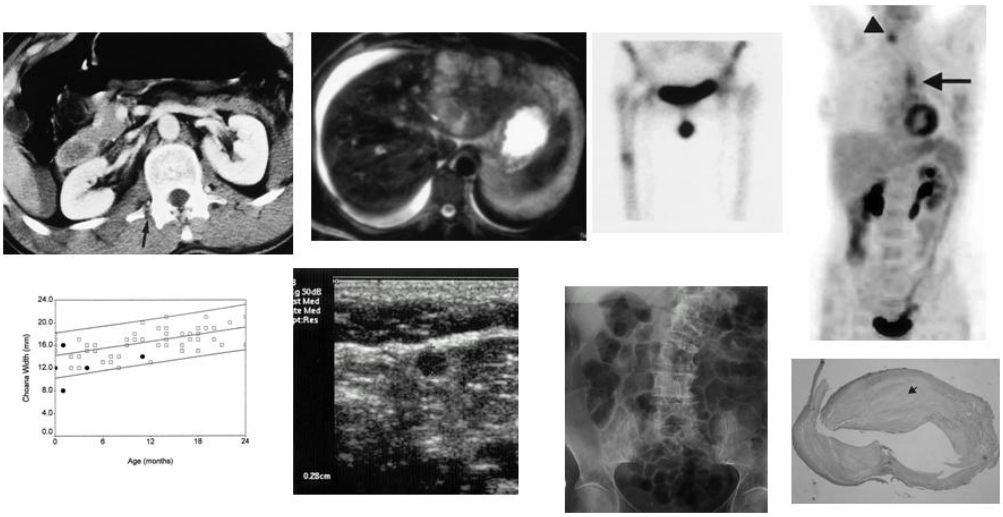
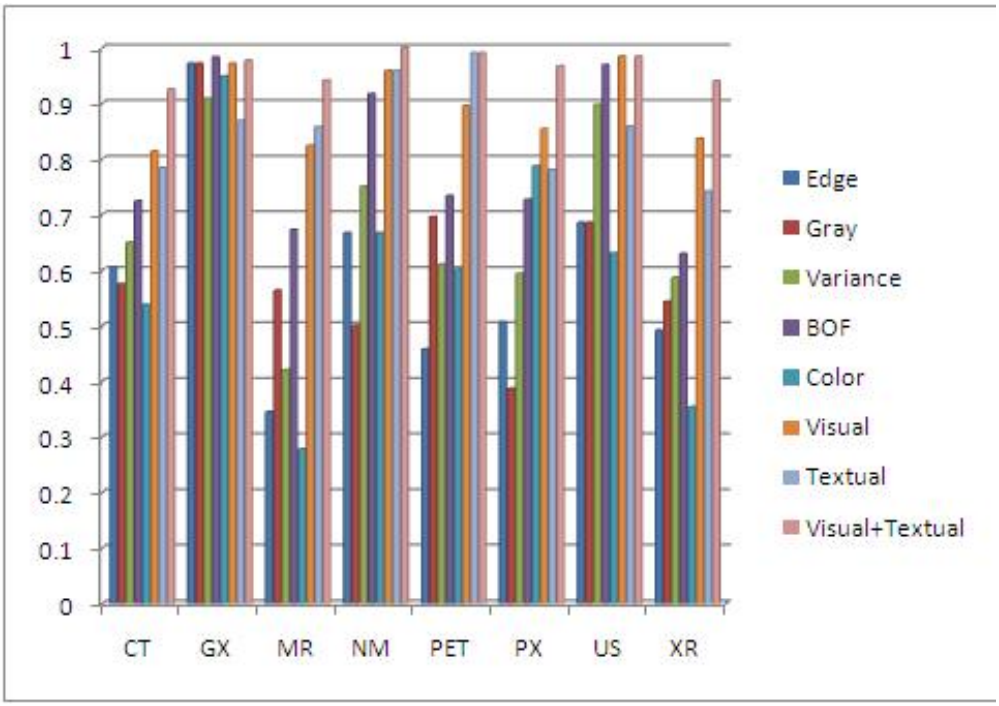
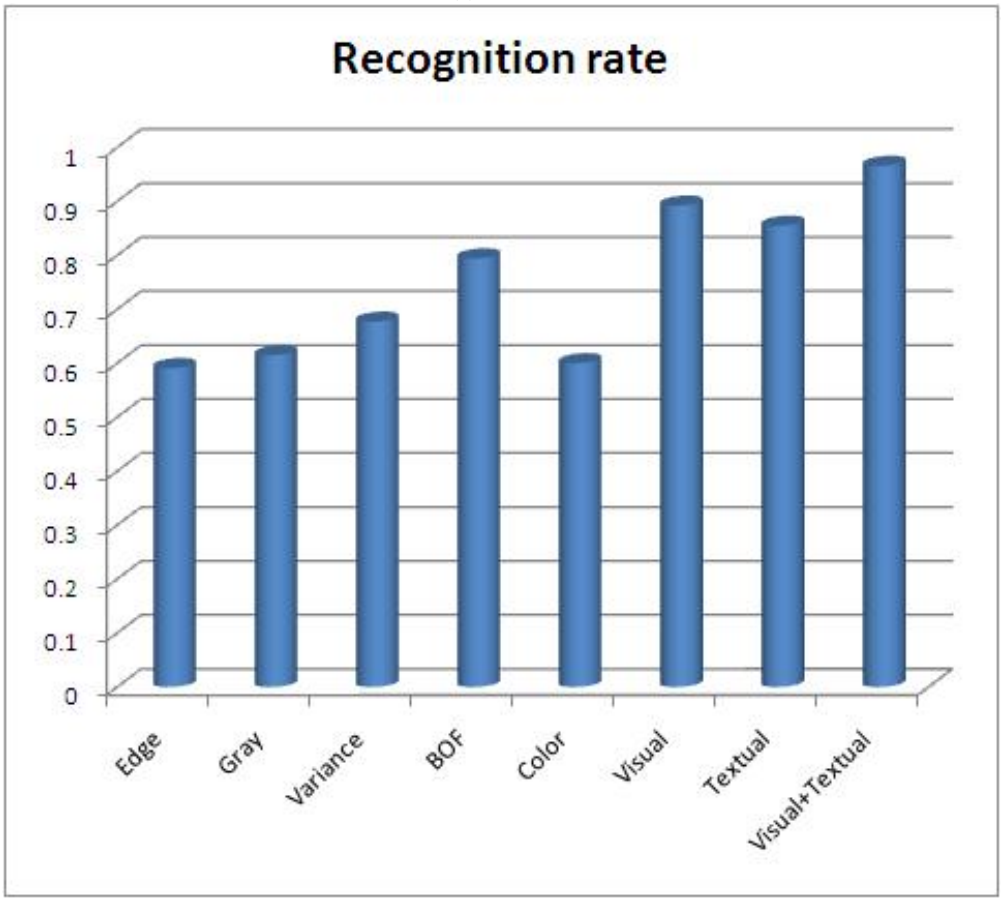


Fig. 1. Sample images of 8 medical modalities. From left to right and upper to down, the images are CT, MR, NM, PET, GX, US, XR and PX, respectively.



(a)



(b)

Fig. 2. The recognition rates of each modality category using different features: Edge, Gray, Color and variance represent the recognition rates of Edge, Gray, Color and variance histogram, respectively; BOF and textual mean the recognition rate of the BOF and textual features; Visual means the recognition rates using combination of all visual features, and textual means those using only textual feature; Visual+texual means those using combination of textual and visual features.

Table 1. Overall calssification rates on medical evaluated dataset using combination of different features.

Features	Visual	Textual	Visual+Texture	Weighted Visual+Textual
classification rate(%)	87.07	84.58	93.36	93.89

4.2 Evaluation

We evaluate classification performance of different features with SVM using training dataset. From each modality set, 180 images are randomly selected for training, the remainder are for test. We did this procedure 5 times, and gave the average classification rate in Fig. 2 (a) for all modality category. From Fig. 2(a), It is obvious that different features have different discriminant for each modality category. For an instance in all visual features, the BOF (Bag-of-Feature) has the best recognition rate for MR modality; however, the color histogram has the best result for PX modality. Therefore, after combining all visual features together, the recognition rate for most of modality can be greatly improved (Fig. 2(a)). The final results with combination of visual and textual feature are also have large improvement that those with only visual or textual feature. Figure. 2(b) gives the average classification rate for all modality. Based on the evaluated better performance with visual feature than textual feature, we also use large weight for visual feature when combine the visual and textual feature.

4.3 Runs Submitted

As Medical Image Processing Group (MIPG) of our Intelligent Image Processing Laboratory (IIPL) In Ritsumeikan University, we prepared four runs for evaluation image set, which used combine visual feature, textual feature, both visual and textual features and weighted visual and textual features. The recognition results are shown in Table 1. We submitted two runs using textual, combined textual and visual features by on-line-system, respectively.

5 Conclusions

In this paper, we proposed to extract different visual and textual features for medical image representation, and fusion the different extracted visual feature and textual feature for modality classification. To extract visual features from the images, we used histogram descriptor of edge, gray or color intensity and block-based variation as global features and SIFT histogram as local feature, and the binary histogram of some predefined vocabulary words for image captions is used for textual feature. Because different features maybe have deferent scale and dimension, in order to allow each individual feature to contribute equally for modality classification, we proposed to use Joint Kernel Equal Contribution (JKEC) for kernel fusion of different features. The proposed algorithm is evaluated by the provided modality dataset by ImageCLEF2010.

ACKNOWLEDGMENT

This work was supported in part by the Grand-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports under the Grand No. 21300070 and 22103513, and in part by the Research fund from Ritsumeikan Global Innovation Research Organization (R-GIRO).

References

1. Muller H, Michoux N, Bandon D, Geissbuhler A., " A review of content-based image retrieval systems in medicine clinical benefits and future directions", International Journal of Medical Informatics 2004, 73, 1-23.
2. Muller H, Deselaers T, Lehmann T, Clough P, Hersh W., "Overview of the ImageCLEFmed 2006 medical retrieval annotation tasks, Evaluation of Multilingual and Multimodal Information Retrieval", Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006, editors: Peters C, Clough P, Gey F, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M, LNCS 2006.
3. Hersh W, Kalpathy-Cramer J, Jensen, J., " Medical Image Retrieval and Automated Annotation: OHSU at ImageCLEF 2006", Working Notes for the CLEF 2006 Workshop, Alicante, Spain. 2006/
4. Guld MO, Kohnen M, Keysers D, Schubert H, Wein BB, Bredno J, Lehmann TM, "Quality of DICOM header information for image categorization", Proceedings SPIE, 4685, 280-287, 2002.
5. Hersh W, Muller H, Jensen J, Yang J, Gorman P, Ruch P., "Advancing biomedical image retrieval: development and analysis of a test collection", J Amer Med Inform Assoc 2006; 13(5).
6. Kalpathy-Cramer J, Hersh W., " Automatic image modality based classification and annotation to improve medical image retrieval", Stud Health Technol Inform. 2007;129(Pt 2):1334-8.
7. Alex Pentland, Rosalind Picard, and Stan Sclaro, "Photobook: Content-based manipulation of image databases", International Journal of Computer Vision, 18(3):233–254, June 1996.
8. Abolfazl Lakdashti and M. Shahram Moin, "A New Content-Based Image Retrieval Approach Based on Pattern Orientation Histogram", Vol. ume 4418, Computer Vision/Computer Graphics Collaboration Techniques,2007.
9. A.K. Jain and A. Vailaya, "Image retrieval using color and shape", Pattern Recognition, Vol. 29, No. 8, pp. 1233-1244, 1996.
10. D. Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, 60(2),pp.91-110, 2004.
11. G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categoraization with bags of keypoints", In Proc. ECCVWorkshop on Statistical Learning in Computer Vision, pp. 1-16,
12. Lazebnik S., Schmid C., Ponce J, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", In Proc. CVPR, pp. 2169- 2178, 2006.
13. Henning Muller, Jayashree Kalpathy-Cramer, Ivan Eggel, Steven Bedrick, Charles E. Kahn Jr., and William Hersh, "Overview of the CLEF 2010 medical image retrieval track", In the Working Notes of CLEF 2010, Padova, Italy, 2010.