A Bag of Semantic Words Model for Medical Content-based Retrieval

[1] Sidong Liu*, [1] Weidong Cai, [1] Yang Song, [2] Sonia Pujol, [2] Ron Kikinis, [1] Dagan Feng

[1] School of Information Technologies, University of Sydney, Australia
[2] Brigham & Women's Hospital, Harvard Medical School, Boston, USA

Abstract. The bag of visual words model has been widely used in content-based image retrieval. However, when it is applied to medical domain, it potentially has several limitations, *e.g.*, some ordinary feature descriptors may not be able to capture the subtle characteristics of medical images; there is a semantic gap between the low-level features and the medical concepts; the emerging multi-modal data pose challenges on current retrieval framework and urge us to extend the possibilities to combine and analyze the multi-modal data. In an attempt to address these issues, we proposed a bag of semantic words model for medical content-based retrieval in this study. We built the high-level semantic features from the low-level visual features by a three-step pipeline. We first extracted a set of low-level features pertaining to the disease symptoms from the medical images. We then translated the low-level features to symptom severity degrees by symptom quantization. Finally, the high-level semantic words were built through learning the patterns of the symptoms. The proposed model was evaluated using 331 multi-modal neuroimaging datasets from the ADNI database. The preliminary results show that the proposed bag of semantic words model could extract the semantic information from medical images and outperformed the state-of-the-art medical content-based retrieval methods.

1 Introduction

Medical imaging technologies, such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), provide important insights for understanding the disease pathology and are essential for biomedical research and health care. However, the increasingly large medical collections pose great challenges in medical data management and retrieval. In addition, there is a wealth of information in medical images, which is important in disease characterization. Therefore, there is a need of medical content-based retrieval (MCBR). MCBR is at the intersection of computer vision, database, and medical informatics, and has been widely used in many applications, such as large data repository management and clinical training and education [1-3]. More importantly, MCBR provides the possibility of accessing a large number of pre-diagnosed patient datasets for clinical decision support.

* Correspondence to S. Liu (sliu7418@uni.sydney.edu.au). This work was supported by ARC, AADRF, NA-MIC (NIH U54EB005149), and NAC (NIH P41RR013218).

Recently, there is a clear trend of using the bag of visual words (BoVW) model for MCBR. Various visual words and models were proposed [4-7]. However, we believe the BoVW models, which are usually based on the low-level features, such as texture, shape, size, intensity, salient points or a combination of them, should be used with caution in MCBR. There are several potential limitations. Firstly, we should note that there is a difference between medical images and ordinary images in that the medical images usually have less variation. When we carry out medical image analysis, we often focus on images acquired on the same structure of interest, such as brain or lung. Many widely used low-level feature descriptors might not be suitable to distinguish the subtle differences between a series of highly similar medical images, as pointed out in [1, 2]. Secondly, there is a semantic gap between the low-level features and the medical concepts. For example, the patients diagnosed with same disease might have dramatically different appearances in medical images based on the low-level visual features. How to associate the low-level features to high-level concepts to bridge the semantic gap in medical image retrieval is yet a problem to be solved. Some studies add semantic annotations [8, 9] to the medical images, but these methods would be less time-effective and prone to errors. Thirdly, the emerging multi-modal medical data could benefit MCBR by providing complementary information, but also pose great challenges on current MCBR framework. It is very challenging to integrate the multi-modal data because they are different in nature. The concatenated features or selected features using feature selection algorithms have been used in many studies [10, 11], *e.g.*, Elastic Net (EN) is one of the state-of-the-art feature selection methods, which not only select the salient single features, but also preserve the correlations between features. However, such feature selection methods are not adding information to the original features; they are actually abandoning the features that they believe are less important. This might lead to biased retrieval due to the loss of information.

In an attempt to address these issues, in this study, we propose a bag of semantic words (BoSW) model for MCBR. We first extract a set of low-level features from the multi-modal medical imaging data and then translate them to the symptom severity degrees by clinical symptom quantization. Finally, we build the high-level semantic words by learning the patterns of the symptoms. We evaluate our method on multi-modal neuroimaging data acquired from the Alzheimer's Disease Neuroimaging Initiatives (ADNI) database. We compare our method to a set of BoVW models that are based on low-level features with and without feature selection [10-12]. The preliminary results show that our proposed BoSW model achieves improved performance compared to other methods.

2 Bag of Semantic Words Model

2.1 Framework Overview

The focus of the BoSW model is to build high-level semantic features from low-level visual features extracted from the medical images. It is a three-step pipeline for deriving the semantic words in the BoSW framework, as shown in Fig. 1.

Fig. 1. The three-step pipeline for deriving the high-level semantic words

In the first step, we need to carefully select the low-level features that could be used to characterize the diseases. Good features are always related to the clinical symptoms. For example, the small volume of a brain functional region in MRI image may indicate atrophy in that region; the low intensity in brain PET image is always interpreted as a sign of neurodegeneration. In the second step, the low-level features are quantitatively associated to the symptoms. In other words, we translate the feature values to the degrees of symptom severity. This step requires applying the knowledge learnt from population-based analyses to the needs of individual patients. Finally, in the third step, we derive the semantic words by learning the patterns of the symptoms for different diseases. This step answers the important question: what are the symptoms of this disease? Note that the symptoms captured by different low-level features can be naturally integrated in this step, because the degrees of symptom severity are directly comparable with each other, unlike the low-level features themselves having different units and different ranges of values. Using this framework, the low-level visual features can be transformed to high-level semantic words and smoothly integrated to the bag of words model.

2.2 Low-level Visual Feature Extraction

In this study, we extract two types of basic low-level visual features from the medical images in two modalities, *i.e.*, the volume fractions from the MRI data to depict the brain atrophy, and the intensity values from PET data to describe the brain degeneration. The volume fraction features ($v_{VOL}^{(i)} \in \mathbb{R}^{N_{ROI}}$) for each subject is defined as the volumes of the N_{ROI} individual brain functional regions normalized by the volume of the entire brain mask, where i indicates the i^{th} subject. The region-wise intensity values ($v_{INT}^{(i)} \in \mathbb{R}^{N_{ROI}}$) for each subject are extracted in the same fashion. Note that $v_{VOL}^{(i)}$ and $v_{INT}^{(i)}$ have the same dimension, because the PET and MRI are registered and segmented using the same brain atlas. More details on MRI and PET data acquisition and pre-processing are given in Section 3.1.

2.3 Clinical Symptom Quantization

In this step, we translate the visual feature values into the degrees of the clinical symptom severity. The underlying assumption for symptom quantization is that the feature values for normal subjects should be randomly distributed a normal range; the extreme feature values might indicate the anomalies.

We used the visual features of a group of normal controls to estimate the normal distribution of the features. The probability distribution function ($P(\mu_j, \sigma_j)$) for each

feature element, $v(j)$, is assumed to be a Gaussian, and its mean (μ_j) and variance (σ_j) can be estimated from the same feature elements of all normal controls, where j indicates the j^{th} lower-level visual feature. We then quantize the value of j^{th} feature for the i^{th} subject using the negative log probability as in Eq. (1):

$$u^{(i)}(j) = -\ln\left(P\left(v^{(i)}(j)|\mu_j, \sigma_j\right)\right) \tag{1}$$

We derive the $u^i(j)$ for all the feature elements and all the subjects. The new feature, $u^{(i)}(j)$, can be interpreted as the possibility of an anomaly. Larger value of $u^{(i)}(j)$ means higher degree of symptom severity.

Originally, the low-level features are not directly comparable to each other, because they have different units and varying ranges of values. After symptom quantization, the feature values are translated to the degrees of symptom severity, which could be naturally compared to each other and combined together.

2.4 Semantic Word Learning

Given the transformed features associated with the symptoms, we then investigate the symptom patterns to derive the high-level semantic words. Usually, a single pattern is derived for each disease showing the dominant symptoms, as in many feature selection methods. However, we notice that even for the same disease, there might be non-unique symptom patterns. For example, Alzheimer's disease (AD) at early stage might cause temporal lobe atrophy, but it could cause frontal and whole brain atrophy at late stage. Based on this fact, we employ the sparse auto-encoder [13] for learning the semantic words, taking advantage of its capability to derive multiple patterns simultaneously. A sparse auto-encoder is a special case of the neural network with three layers, the input layer, hidden layer and output layer, as shown in Fig. 2. Different from a typical neural network, the goal of a sparse auto-encoder is to learn the internal structures of the input features instead of predicting the classes of them. The neurons are constrained to output the same piece of information as in the input features. In this study this is equivalent to optimally identifying the most possible combinations of symptoms of the disease.

The weights of the neurons, W, could be estimated by energy minimization of three types of cost, *i.e.*, the Error Cost, Weight Cost and Sparsity Cost, as in Eq. (2):

$$\arg\min_{W} \overbrace{\left[\frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{2}\|\hat{u}^{(i)} - u^{(i)}\|^2\right)\right]}^{Error\ Cost} + \overbrace{\frac{\lambda}{2}\sum_{l=1}^{2}\sum_{k}^{N_I \times N_H} W(k)_l^2}^{Weight\ Cost} + \overbrace{\beta\sum_{h=1}^{N}\mathrm{KL}(\rho\|\hat{\rho}_h)}^{Sparsity\ Cost} \tag{2}$$

where M is the number of subjects, $\hat{u}^{(i)}$ is the estimated output of $u^{(i)}$; N_I and N_H refer to the numbers of input neurons and hidden neurons, respectively; W_1 and W_2 are $N_I \times N_H$ and $N_H \times N_I$ matrices representing the weights on the neurons in conjunctive layers; $\hat{\rho}_h$ is the average activation of h^{th} hidden neuron; $\mathrm{KL}(*\|*)$ is the Kullback-Leibler divergence between two variables. We could use λ, β and ρ to control the ratios of the three cost functions. Each neuron in the hidden layer shows a combination of signal received from the input neurons with the maximum activation value. In other words, the hidden neurons show the patterns of the most possible symptoms for the

disease. Note that sparse auto-encoder provides us the flexibility to define arbitrary number of hidden neurons (N_H) to capture the non-unique symptom patterns.

In this study, we divide the subjects into different groups according to their diagnosed diseases and investigate each group individually to identify the disease-specific symptoms. Fig. 2 shows the architecture of the sparse auto-encoder for the three groups of subjects. More details of the subject information are given in Section 3.1.

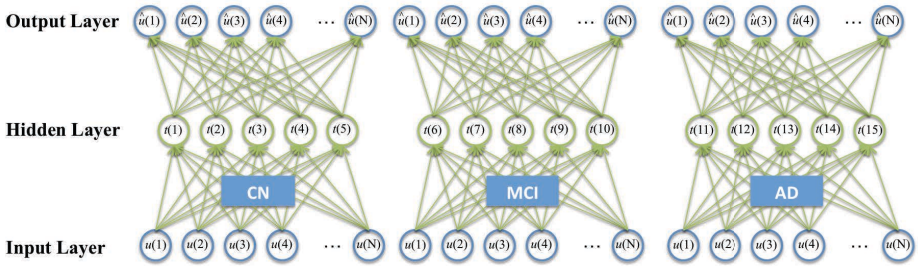

Fig. 2. Architecture of sparse auto-encoder used in this study, designed for different diseases

We use the hidden neurons derived from all these groups as the final collection of semantic words. The semantic features for each subject are derived by calculating the activation values to each semantic word, given his/her symptom severity features. In this study, we use the sigmoid function for calculating the activation values. The low-level features for each subject are then transformed to a set of activation values to the semantic words, which we interpret as the 'term frequencies'; therefore, the high-level semantic features can be smoothly integrated to the typical bag of word model for ordinary information retrieval systems.

3 Experiments

3.1 Data Acquisition and Pre-processing

The medical imaging data used in this study were obtained from the ADNI database [14]. We randomly selected 331 subjects from the ADNI baseline cohort, including 77 cognitive normal (CN) subjects, 169 Mild Cognitive Impairment (MCI) and 85 Alzheimer's Disease (AD). For each subject, we acquired a T1-weighted volume on a 1.5 Tesla MR scanner and a PET volume with Fluorodeoxyglucose (^{18}F) as the tracer. All these 3D MRI and PET data were converted to the ADNI format following the ADNI image correction protocols [14, 15]. The PET images were linearly registered to the corresponding MRI image using FSL FLIRT [16]. We further nonlinearly registered the MRI images to the ICBM_152 template [17], which parcellated the brain into 83 brain functional regions, using the Image Registration Toolkit (IRTK) [18]. We then applied the IRTK registration coefficients to warp the linearly registered PET images into the ICBM_152 template. All of the brain functional regions in registered PET and MRI were labeled in the template space using the multi-atlas propagation with enhanced registration (MAPER) approach [19].

3.2 Performance Evaluation

Our proposed BoSW model was validated by leave-one-out cross-validation on the entire dataset using the query by example paradigm. The similarity between any two feature-vectors was calculated using the normalized Euclidean distance. We evaluated its performance using the revised version of the mean average precision (MAP), same as in [11].

The number of subjects in different groups was based on the subjects' diagnoses ($[M_{CN}, M_{MCI}, M_{AD}] = [77, 169, 85]$). The number of regions was determined by the ICBM_152 template ($N_{ROI} = 83$), thus the number of input neurons was equal to the sum of both PET and MRI features ($N_I = 166$). We assumed the symptom pattern associated to each disease might not be unique, so we set a number of hidden neurons to capture the multiple patterns simultaneously ($N_H = 5$) for each group of subjects; therefore there were 15 hidden neurons in total for all three groups. Other parameters of sparse auto-encoder were set by pilot experiments ($[\lambda, \beta, \rho] = [0.0001, 3, 0.01]$). We were not dedicated to tune the parameters to achieve the best performance in this study. This preliminary experiment was simply designed to demonstrate the superiority of the BoSW model.

We compared the diagnosis performance of the proposed BoSW model to a set of BoVW models that were based on the PET features [12], MRI features [11], and concatenated PET + MRI features with and without feature selection by Elastic Net [10]. Same performance evaluation methods were used for all these models.

3.3 Results

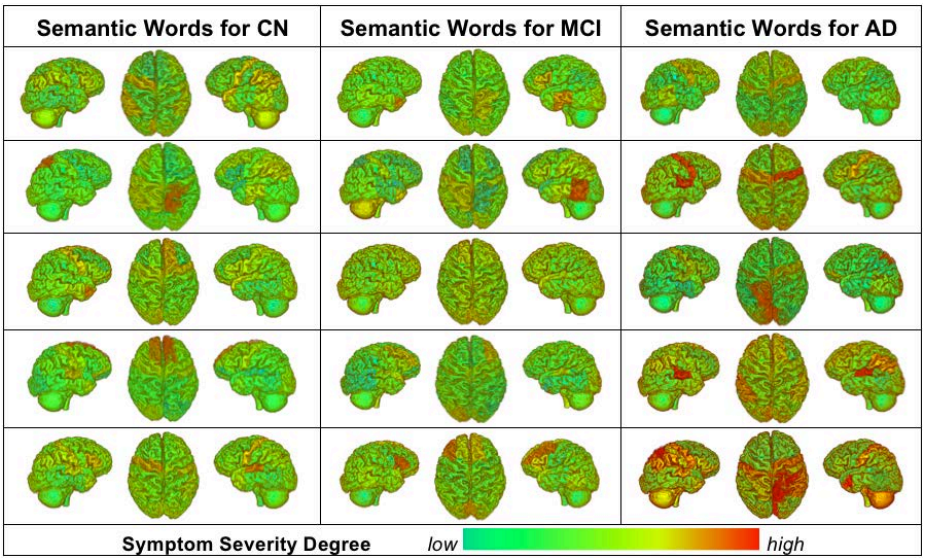

Fig. 3. The back-projection of the semantic words onto the brain, image generated using 3D Slicer [20]

The semantic words derived by our BoSW pipeline are shown in Fig. 3. Each semantic word is derived from a hidden neuron using the sparse auto-encoder as discussed

in Section 2.4. A color table is used to illustrate the symptom severity degrees combining both the MRI and PET components. It is easy to interpret the semantic words using simple symptom descriptions. For example, the last neuron derived for AD shows that an AD patient may have strong correlations between the temporal and parietal lobes with both atrophy and hypo-metabolism symptoms.

Table 1. The mean average precision (%) of the proposed BoSW model compared to the state-of-the-art BoVW models based on low-level visual features

Diagnosis / Features (Dimension)	CN	MCI	AD	Overall
PET Features (83)	50.4	75.3	48.2	62.6
MRI Features (83)	59.2	69.2	53.8	62.4
Concatenated Features (166)	55.0	72.5	51.3	62.9
EN Selected Features (23)	66.9	69.7	58.5	66.2
The Proposed Semantic Words (15)	58.2	76.1	54.5	66.4

Table 1 shows the performance comparison of the proposed BoSW model and a set of BoVW models that are based on low-level visual features. The results correspond well to the existing knowledge of AD and MCI. PET features are more sensitive to early functional changes on MCI, so they have better retrieval performance for MCI. MRI features are more robust to structural changes that are mainly found in AD. The concatenated PET + MRI features could not taking the advantages of multi-modal information, and show a moderate performance compared to PET and MRI features. The selected features using Elastic Net yield best results for NC and AD at a sacrifice of compromised performance for MCI. This indicates that feature selection methods might lead to biased retrieval due to loss of information. The proposed BoSW model on the other hand, preserves all the goodness of individual features and achieved the best overall performance and best performance for MCI. The advantageous multi-pattern nature of sparse auto-encoder makes the BoSW model more robust than the feature selection algorithms. Note that our BoSW model uses as few as 15 words, and the performance can be further improved through tuning the parameters or using different low-level features. More importantly, our BoSW model could make sense of the low-level features and attempt to bridge the semantic gap by translating them to the high-level semantic concepts.

4 Conclusions

In this study, a novel bag of semantic words (BoSW) model and its application in MCBR were presented. Our BoSW model used much fewer features and could capture the high-level semantic words highly correlated to disease symptoms. Evaluated using a public multi-modal neuroimaging dataset, our method outperformed other state-of-the-art MCBR methods.

References

1. Muller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medical applications – clinical benefits and future directions. International Journal of Medical Informatics 73, 1-23. (2004)
2. Cai, W., Kim, J, Feng, D.: Chaper 4 – content-based medical image retrieval. In: Feng, D. (eds.) Biomedical Information Technology, pp. 83-113. Elsevier (2008)
3. Long, L.R., Antani, S., Deserno, T.M., Thoma, G.R.: Content based image retrieval in medicine. Int. J. Healthc. Inf. Syst. Inform. 4(1), 1-16. (2009)
4. Liu, S., Cai, W., Wen, L., et al.: Localized functional neuroimaging retrieval using 3D discrete curvelet transform. In: ISBI 2011, pp. 1877-1880. IEEE (2011)
5. Burner, A., Donner, R., Mayerhoefer, M., et al.: Texture bags: anomaly retrieval in the medical images based on local 3D-texture similarity. In: Muller, H., Greenspan, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2011. LNCS, vol. 7075, pp. 116-127. Springer, Heidelberg (2011)
6. Foncubieta-Rodriguez, A., Depeursinge, A., Muller, H.: Using multiscale visual words for lung texture classification and retrieval. In: MCBR-CDS 2011. LNCS, vol. 7075, pp. 69-79 (2011)
7. Haas, S., Donner, R., Burner, A., Holzer, M., Langs, G.: Superpixel-based interest points for effective bags of visual words medical image retrieval. In: MCBR-CDS 2011. LNCS, vol. 7075, pp. 58-68. Springer, Heidelberg (2011)
8. Moller, M., Sintek, M.: A generic frameword for semantic medical image retrieval. In: KAMC 2007, vol. 253, no. 2. CEUR (2007)
9. Seifert, S., Thoma, M., et al.: Combined semantic and similarity search in medical image databases. In: Boonn, W.W., Liu, B.J. (eds.) SPIE-MI 2011, vol.7967, no.2. SPIE (2011)
10. Shen, L., Kim, S., Qi, Y., Inlow, M., Swaminathan. S., Nho, K., Wan, J., et al.: Identifying neuroimaging and proteomic biomarkers for MCI and AD via the elastic net. In: Liu, T., Shen, D., et al. (eds.) MBIA 2011. LNCS, vol. 7012, pp. 27-34. Springer, Heidelberg (2011).
11. Liu, S., Cai, W., Wen, L., Feng, D.: Multi-channel brain atrophy analysis in neuroimaging retrieval. In: ISBI 2013, pp. 206-209. IEEE (2013)
12. Cai, W., Liu, S., Wen, L., et al.: 3D neurological image retrieval with localized pathology centric CMRGlc patterns. In: ICIP 2010, pp. 3201-3204. IEEE (2010)
13. Raina, R., Battle, A., Lee, H., et al.: Self-taught learning: transfer learning from unlabeled data. In: The 24th International Conference on Machine Learning, pp. 759-766. ACM (2007)
14. Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., et al.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. JMRI 27(4), 685–691 (2008)
15. Jagust, W.J., Bandy, D., Chen, K., Foster, N.L., et al.: The Alzheimer's disease neuroimaging initiative positron emission tomography core. Alzheimer's & Dementia 6(3), 221-229 (2010)
16. Jenkinson, M., Bannister, P., et al.: Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17(2), 825-841 (2002)
17. Mazziotta, J., Toga, A., Evans, A., Fox, P., et al.: A probabilistic atlas and reference system for the human brain: international consortium for brain mapping. Phil. Trans. Royal Soc. B Biol. Sci. 356(1412), 1293-1322 (2001)
18. Schnabel, J.A., Rueckert, D., Quist, M., et al.: A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In Niessen, W.J., Viergever, M.A. (eds.) MICCAI 2001. LNCS, vol. 2208, pp. 573-581. Springer, Heidelberg (2001)
19. Heckemann, R.A., Keihaninejad, S., Aljabar, P., Gray, K.R., et al.: Automatic morphometry in Alzheimer's disease and mild cognitive impairment. Neuroimage 56(4), 2024-2037 (2011)
20. Pieper, S., et al.: The NA-MIC kit: ITK, VTK, pipelines, grids and 3D Slicer as an open platform for the medical image computing community. In: ISBI 2006, pp. 698–701. IEEE (2006)