Multimodal Medical Image Retrieval based on Latent Topic Modeling

Mandikal Vikram, Aditya Anantharaman, Suhas BS, Sowmya Kamath S
Email: {15it217.vikram, 15it201.aditya.a, 15it110.suhas, sowmyakamath}@nitk.edu.in
National Institute of Technology Karnataka, Surathkal

Abstract

Modern medical practices are increasingly dependent on Medical Imaging for clinical analysis and diagnosis of patient illnesses. A significant challenge when dealing with the extensively available medical data is that it often consists of heterogeneous modalities. Existing works in the field of Content based medical image retrieval (CBMIR) have several limitations as they focus mainly on visual or textual features for retrieval. Given the unique manifold of medical data, we seek to leverage both the visual and textual modalities to improve the image retrieval. We propose a Latent Dirichlet Allocation (LDA) based technique for encoding the visual features and show that these features effectively model the medical images. We explore early fusion and late fusion techniques to combine these visual features with the textual features. The proposed late fusion technique achieved a higher mAP than the state-of-the-art on the ImageCLEF 2009 dataset, underscoring its suitability for effective multimodal medical image retrieval.

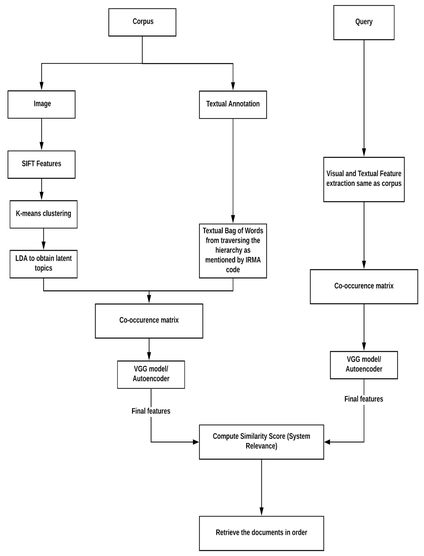
Visual feature extraction

The Scale-invariant Feature Transform is used to detect the salient points and obtain the feature descriptors of a given image. The feature descriptors of all the images in the corpus are clustered into some k clusters using the K-means algorithm. The hence obtained k centroids are referred to as the “visual words”. All the SIFT descriptors of an image are represented by their respective nearest visual words. The images are modeled as a histogram of visual words to represent the corpus as a bag of visual words. We adopt an approach where the images are represented as a histogram of latent topics. Latent topics are derived from the visual words using the Latent Dirichlet Allocation (LDA) and are dubbed as the “visual topics”.

Textual feature extraction

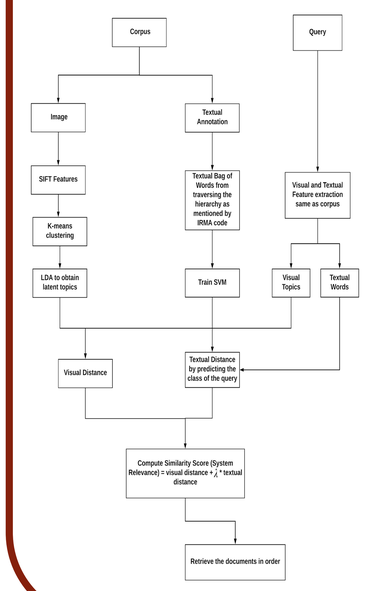
Each image in the ImageCLEF 2009 dataset is represented by its IRMA (Image Retrieval in Medical Application) code. The code has four independent axes, each of which describes a different aspect of the images. The syntax of the IRMA code is of the form *TTTT-DDD-AAA-BBB*, we extract the words from the last two axes (A and B) which describe the region of the body and the biological system to which the part in the image belongs. This is done by traversing down the hierarchy as specified by the code and recording all the words along this traversal. This process is performed independently for both A and B, and all words along both the traversals as represented as textual words.

Early Fusion



- Co-occurrence matrix is calculated for all the textual and visual words.
- VGG based model or autoencoder used for compressing the co-occurrence matrix.
- The distance is given by $\mathcal{D}_e(x_i, q) = \|\mathcal{F}_e(x_i) - \mathcal{F}_e(q)\|_2$ where $\mathrm{F}_e(d)$ denotes the compressed early fusion features of the document d.

Late Fusion

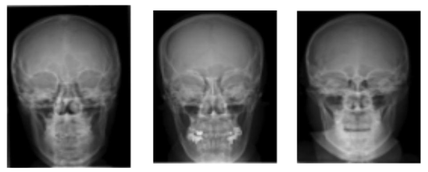


- Visual distance is given by $\mathcal{D}_v(x_i, q) = \|\mathcal{F}_v(x_i) - \mathcal{F}_v(q)\|_2$ where $\mathrm{F}_v(d)$ is the latent visual features of document d
- Textual distance is given by $\mathcal{D}_t(x_i, q) = (\mathcal{C}(x_i) - \mathcal{C}(q))^2$ where $\mathcal{C}(d)$ is the class predicted for the document d by the text-feature based SVM
- Final distance $\mathcal{D}_l(x_i, q) = \mathcal{D}_v(x_i, q) + \lambda \mathcal{D}_t(x_i, q)$

Results

Text: skull

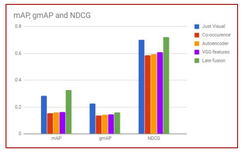
QUERY



Top retrieved results using late fusion

mAP, gmAP and NDCG

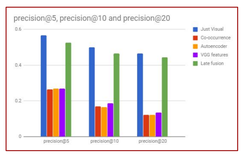
Approach	mAP	gmAP	NDCG
Visual features only	0.283	**0.226**	0.702
Co-occurrence matrix	0.155	0.138	0.587
Autoencoder features	0.16	0.143	0.596
VGG features	0.162	0.145	0.61
Late fusion	**0.326**	0.159	**0.722**



Previous best mAP: 0.283

precision@5, precision@10, precision@20

Model	p@5	p@10	p@20
Visual features only	0.567	0.5	0.465
Co-occurrence matrix	0.264	0.169	0.122
Autoencoder features	0.268	0.165	0.123
VGG features	0.27	0.187	0.134
Late fusion	0.526	0.466	0.445



Conclusion

In this we propose a LDA based technique to encode the visual features of medical images and explore various techniques to fuse the visual and textual modalities. Our late fusion techniques outperforms the previous state of the art on the ImageCLEF2009 dataset. In view of this, we intend to further explore the semantic relationships between textual and visual words, so that the proposed fusion techniques could be improved. The code can be found on https://github.com/vikram-mm/Multimodal-Image-Retrieval. We are also working on extending our late fusion approach so that it can be applied to larger corpora.

We gratefully acknowledge the funding and computing facilities provided by Science and Engineering Research Board, Department of Science & Technology, Government of India, as part of the Early Career Research Grant (Grant no: ECR/2017/001056) to the fourth author.