An Approach for Multimodal Medical Image Retrieval using Latent Dirichlet Allocation

Mandikal Vikram
Department of Information Technology,
National Institute of Technology Karnataka, Surathkal
15it217.vikram@nitk.edu.in

Aditya Anantharaman
Department of Information Technology,
National Institute of Technology Karnataka, Surathkal
15it201.aditya.a@nitk.edu.in

Suhas BS
Department of Information Technology,
National Institute of Technology Karnataka, Surathkal
15it110.suhas@nitk.edu.in

Sowmya Kamath S
Department of Information Technology,
National Institute of Technology Karnataka, Surathkal
sowmyakamath@nitk.edu.in

Abstract

Modern medical practices are increasingly dependent on Medical Imaging for clinical analysis and diagnoses 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.

CCS Concepts • **Information systems → Multimedia and multimodal retrieval**; **Image search**; *Novelty in information retrieval*; Information extraction; • **Applied computing → Health care information systems**;

Keywords Content-based Image Retrieval (CBIR), Topic modeling, Medical Informatics

CoDS-COMAD '19, January 3–5, 2019, Kolkata, India

ACM ISBN 978-1-4503-6207-8/19/01…$15.00
https://doi.org/10.1145/3297001.3297007

ACM Reference Format:
Mandikal Vikram, Aditya Anantharaman, Suhas BS, and Sowmya Kamath S. 2019. An Approach for Multimodal Medical Image Retrieval using Latent Dirichlet Allocation. In *6th ACM IKDD CoDS and 24th COMAD (CoDS-COMAD '19), January 3–5, 2019, Kolkata, India.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3297001.3297007

1 Introduction

Medical Imaging provides a wide variety of techniques for creating visual representations of the internal organs of the human body for diagnostic and clinical purposes. These are progressively becoming integral and indispensable constituents of medical intervention and analysis. The proliferation in medical image data from medical institutions, documented in digital forms is a significant valuable asset for diagnostic medical informatics. This data provides an invaluable source of information for diagnostic studies related to diseases like cancer, tumors, fractures etc. Medical Image retrieval can serve as an useful tool for medical practitioners and help them effectively utilize the large amount of digital medical data available in making clinical decisions. The primary objective of such medical image retrieval tasks is to retrieve images which are the most relevant from a given clinical perspective [2]. Several text based retrieval techniques have been proposed over the years, but, their significant drawback remains their high dependency on textual annotations for the images. Such textual annotation are often incomplete, ambiguous or completely absent.

This varying subjectivity and shallow context sensitivity of the image annotations are significant hurdles faced by text retrieval techniques. Hence, more recent approaches have focused more on combining visual and textual techniques for a multi-modal approach to image retrieval. One of the major challenges in medical image retrieval is that the low-level visual and textual features do not directly correspond to the high-level medical concepts, in other words, there exists a semantic gap between the two. Another challenge is the manner in which the visual and textual features are integrated,

which needs to be specifically addressed to account for the manifold information contained in medical images.

We propose the use of latent topics generated for the visual features of medical images for effective medical image retrieval. We propose a Latent Dirichlet Allocation (LDA) based technique for visual feature extraction in which the latent topics derived using the LDA[1] from the generated SIFT features[9] are used as the visual features. To the best of our knowledge, such an approach has not been used in medical image retrieval. We then propose various fusion approaches for combining the visual and textual features and capture the correlation between them. We make the source code available publicly in a GitHub repository[1].

2 Related Work

Existing techniques for Medical Image Retrieval can be categorized into three approaches - Content Based Medical Image Retrieval (CBMIR), multi-modal fusion based image retrieval and deep learning based image retrieval. CBMIR based approaches focus on adapting the concepts of Content Based Image Retrieval (CBIR) for medical images, and has been a field of active research in the recent years. In CBIR, the main goal is to organize images based on their visual content and features.

Greenspan et al [4] developed a continuous and probabilistic image representation scheme using Gaussian mixture modeling (GMM) with information-theoretic image matching based on the Kullback-Leibler (KL) measure. However, GMM is a very crude (and lossy) image representation, which is a significant drawback of their method. Also, the number of Gaussians (k) is defined and kept constant per image, which affects the classification accuracy. Rahman et al [14] used a novel multimodal query expansion to design a framework which integrates visual and textual keywords. However, because of the low-level continuous feature representation used in most CBIR systems, the idea of query expansion cannot be directly applied to existing systems. Another system for image retrieval using radiological images was proposed by Napel et al [10] that used a dataset of CT images annotated by radiologists for retrieving similar lesions in scan images.

Fusion based approaches aim to bridge the gap between the semantics of different modalities using information fusion techniques. They can be further classified into two categories - feature fusion and retrieval fusion. Feature fusion approaches involve the generation of an integrated feature representation from many modalities. Approaches such as [11] focus on the generation of feature vectors by simply concatenating the normalized features from different modalities. More structured approaches such as [8] proposed a multilayer PLSA (Probabilistic Latent Semantic Analysis)[5] based model for the problem of image retrieval using multiple modalities. Cao et al [2] used a PLSA based approach to bridge the semantic gap between the textual and visual modalities. Retrieval fusion involves merging of retrieval results from many retrieval algorithms to generate the results. Tai et al [16] proposed a method for image retrieval based on fuzzy c-means using spatial weighted entropy. Their approach did not take into account relationships between shape, texture and other high-level information, due to which performance suffered. Huang et al [6] proposed a query dependent feature fusion method for medical image retrieval based on one class SVM.

Recent advancements in the field of deep learning have resulted in many approaches for medical image retrieval that can learn representations of images and textual features for better performance tuning. Qayyum et al [13] proposed a deep CNN based model for classifying medical images, which is then used for medical image retrieval. However, they generated a limited dataset for testing the model which was not diverse enough to be adapted for 3D volumetric applications. Pyykko et al [12] proposed a deep learning framework that utilizes pre-extracted features from CNNs and learns a new distance representation based on the users' relevance feedback. Although this model takes less time for training it relies a lot on user feedback, which is often very difficult to obtain. Our approach attempts to overcome the problems associated with all these models, and we also show why a LDA based model has certain additional advantages over these methods.

3 Proposed Approach

Figure 1 depicts the overall workflow of the proposed approach. Each of these phases are described in detail next.

3.1 Visual Feature Generation using LDA

We used the standard ImageCLEF 2009 dataset [7] for the experimental validation of the proposed methodology. We adopt a Visual Bag-of-Words (VBoW) model inspired by [2] to represent the visual features. The Scale-invariant Feature Transform [9] 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 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, i.e. the cluster center of their respective cluster. The images are modeled as a histogram of visual words to represent the corpus as a bag of visual words. In contrast to the classic VBoW model, 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) [1] and are dubbed as the "*visual topics*". We observed that words mapped to a given visual topic are semantically closer. Hence, it would be more efficient to represent a document using these few

[1] https://github.com/vikram-mm/Multimodal-Image-Retrieval

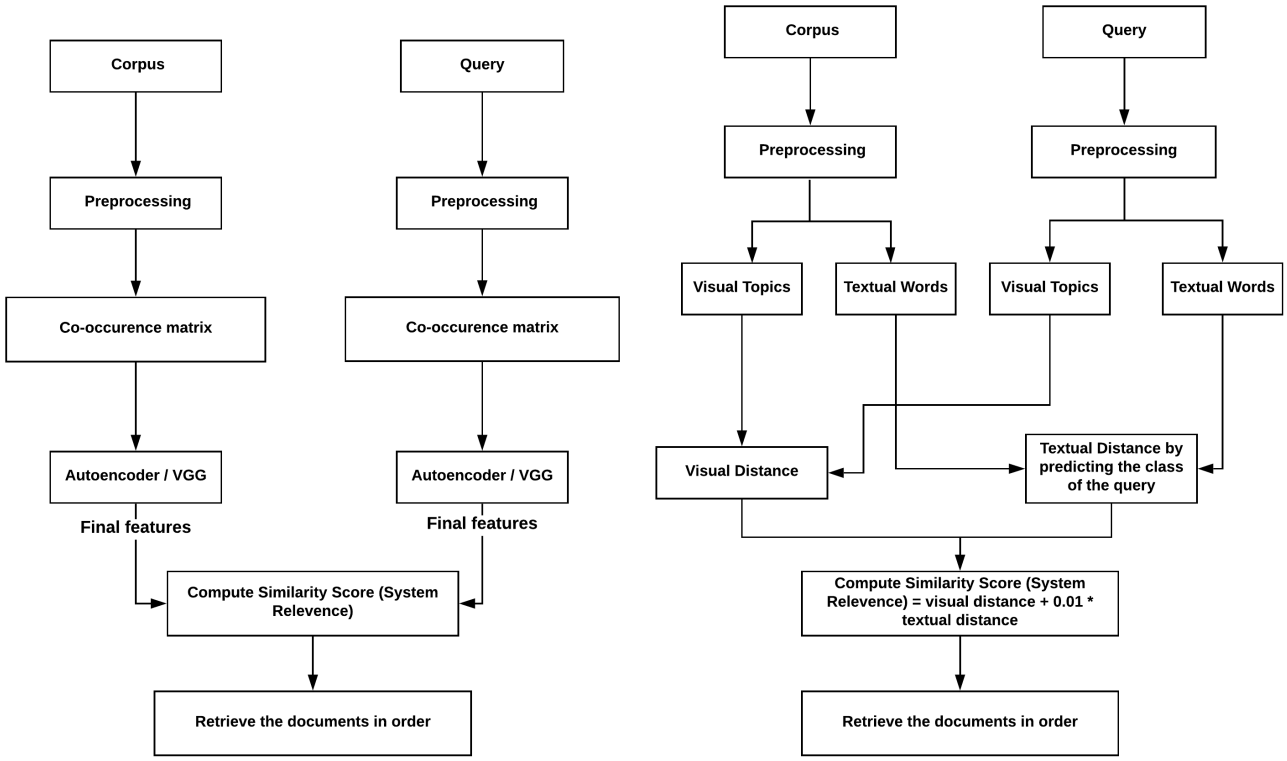


Figure 1. Proposed Visual Topic Modeling based Approach for Multi-modal Medical Image Retrieval

visual topics than using the visual words. We also found that the proposed LDA based approach exceptionally models the visual features in medical images and is a lot more effective than the PLSA based latent topic modeling used in [2]. This can be attributed to the fact that the Dirichlet prior on the per-document distribution prevents the well-known overfitting in PLSA to a large extent.

During LDA modeling, the visual words of an image are grouped into topics which represent the image and each word can be attributable to a topic. Figure 2 gives an overview of the parameters of an LDA model and how it works. θ is the topic distribution for a document m, α is the Dirichlet prior on the topic distributions in a document, β is the parameter of the Dirichlet prior on the word distribution in a topic, z_{mn} is the topic for the n^{th} word in document m, w_{mn} is the specific word, and ϕ is the word distribution for topic k. The visual words are grouped into latent topics by the LDA model which give the visual topics. Semantically closer words are mapped to the same visual topic and hence a given document (in our case, medical images) can be represented in terms of visual topics. These visual topics are our final visual features.

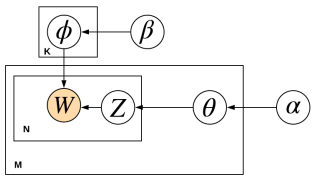


Figure 2. LDA Model parameters

3.2 Textual Feature Generation

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*, where the technical factor T represents the image modality, D gives the direction of the body orientation, A describes the region of the body examined while B denotes the functional biological system described. 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. We propose techniques based on early and late fusion concepts for combining both the visual and textual modalities, which we describe in the next section.

3.3 Early Fusion Approaches

Fig. 1(a) depicts the overall early fusion process designed for retrieving the best matched images. This involves combining the visual and textual features by computing a co-occurrence matrix of the visual topics and the associated textual, this co-occurrence matrix is calculated for each of the images. As this co-occurrence matrix is most often sparse, it cannot be used as a feature on its own. Two different models are employed to extract relevant features from this co-occurrence matrix which are described in detail next.

The first model is an autoencoder model which compresses the co-occurrence matrix to a smaller matrix before retrieving similar images. An autoencoder is a neural network that has three layers: an input layer, a hidden (encoding) layer, and a decoding layer. The network is trained to reconstruct its inputs, which forces the hidden layer to try to learn good representations of the inputs. The key aspect of an autoencoder is the informational bottleneck which forces the network to form an intermediate representation of smaller dimensionality than that of the input. Due to this, the network is constrained to retain only those components which can be used to reconstruct the common features, while rejecting irrelevant features. Thus, we use the output of the bottleneck layer of the trained autoencoder as the compact feature representation of the inputs.

The second model we used for early fusion is a pre-trained variant of the popular VGG-16 model [15]. The VGG-16 model is an image classification CNN which performed exceptionally well in the ImageNet challenge[3]. Since the classical VGG network has been extensively trained over a vast number and variety of images (none of which are medical images though), the trained layers have attained a high learning rate towards capturing relevant features from a given image input. Thus, we adapt the classical VGG-16 model, and use only some of its layers to extract relevant features from the co-occurrence matrix for our image dataset. In our model, only the first 4 layers of the VGG-16 model were used to obtain a compressed representation of the co-occurrence matrix before retrieving similar images.

After obtaining the features from the autoencoder model or the VGG-16 based model, these are used for retrieval and the retrieval performance is evaluated. The image similarity is computed based on the Euclidean distance between the image features, as given by the compressed representation from the autoencoder or VGG models. Let x_i denote the i^{th} document in the corpus and let q denote the query, then their distance (or dis-similarity) in the early fusion approach $\mathcal{D}_e(x_i, q)$ is given by

$$\mathcal{D}_e(x_i, q) = \|\mathcal{F}_e(x_i) - \mathcal{F}_e(q)\|_2 \tag{1}$$

where $\mathcal{F}_e(d)$ denotes the compressed early fusion features of the document d. For a given query image, those top-k images which have the smallest Euclidean distance value are retrieved from the dataset.

3.4 Late Fusion Approach (Ensemble Model)

An ensemble model that builds on both the visual and textual features is proposed and incorporated as a late fusion technique for determining best matching images. The visual and textual features are generated as described earlier in Section III(B). For a query q and document x_i, the visual distance is $\mathcal{D}_v(x_i, q)$ is defined as -

$$\mathcal{D}_v(x_i, q) = \|\mathcal{F}_v(x_i) - \mathcal{F}_v(q)\|_2 \tag{2}$$

where $\mathcal{F}_v(d)$ denotes visual feature of the document d i.e. the latent topics of document d. The ImageCLEF dataset also provides class information for each image. When one image is queried, all the retrieved images belonging to the same class can be considered to be relevant while the ones belonging to different classes can be discarded as irrelevant. We trained a Support Vector Machine (SVM) to predict the class of the image when presented with just its textual features. Hence, the textual distance $\mathcal{D}_t(x_i, q)$ given by,

$$\mathcal{D}_t(x_i, q) = (C(x_i) - C(q))^2 \tag{3}$$

where $C(d)$ denotes the class predicted for the document d by the text-feature based SVM. We define the textual loss as in Equation (3) since labels which are closer to each other represent semantically similar classes in the ImageCLEF 2009 datatset. Thus, the total distance $\mathcal{D}_l$ for a document x_i and query q in late fusion approach $\mathcal{D}_l(x_i, q)$ is given by the combination of the visual distance in Equation (2) and the textual distance in equation 3, which is,

$$\mathcal{D}_l(x_i, q) = \mathcal{D}_v(x_i, q) + \lambda \mathcal{D}_t(x_i, q) \tag{4}$$

The value of λ in Equation (4) was heuristically determined to be 0.01. The small value of λ indicates that the coarser ranking is done by the visual distance, while, the textual distance is used for finer re-ranking. Fig. 1(b) depicts the late fusion process in detail.

4 Experimental Results & Evaluation

We used the ImageCLEF 2009 dataset, a standard and open dataset used widely by several state-of-the-art works, for validating the proposed medical image retrieval task. The dataset contains 12669 medical images from both radiology and radio-graphical origins. To these images, we first applied the visual feature extraction.

The visual features were initially modeled as a visual bag-of-words (VBoW). The SIFT features were then extracted and clustered into 3000 clusters. Due to the large number

of SIFT features obtained, it was infeasible to implement the clustering sequentially, hence we implemented on the GPU environment using TensorFlow. Each image was then represented by the set of cluster numbers of all the SIFT features present which comprise the VBoW vocabulary. The SIFT key points obtained are shown in Fig. 3, and each of the corresponding SIFT features are of length 128.

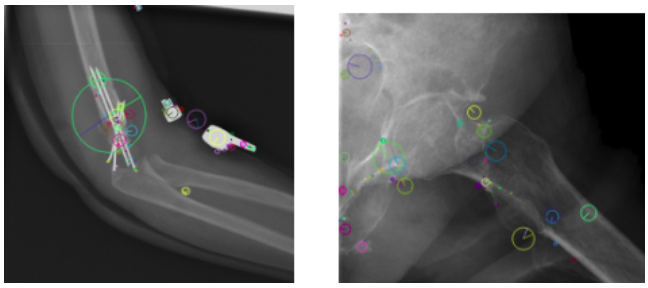


Figure 3. Extracted SIFT key points

On this VBoW vocabulary, LDA is performed to obtain 100 latent topics. The visual features are now represented by the probability of each of the 100 topics given by the LDA model. Evaluation was performed directly using the visual features modeled in this way, and our experiments showed that retrieval results outperformed succeeding experiments involving only textual features. The textual features were simply the one hot feature vectors of vocabulary size based on the keywords present in the image caption. The textual vocabulary size was 272 when compared to the visual vocabulary (number of visual topics) of 100.

Next, a co-occurrence matrix was created of dimensions $textual_{vocab} \times visual_{vocab}$. The probabilities obtained from the LDA model are populated in the matrix in those rows where each particular textual word is present in the caption. Every image is hence represented by such a co-occurrence matrix and various retrieval experiments were performed using these co-occurrence matrices. We observed a sharp drop in performance, which indicates that the high sparsity of the representation is the culprit.

The next logical step was to reduce this sparsity and to this end, we employed the autoencoder based early fusion model. The 272×100 co-occurrence matrix was now reduced to 34×13. These dense features were then used for retrieval. The evaluation parameters marginally improved but were still a long way short of what was obtained just using the visual features. The next experiment was to use the VGG features extracted from the co-occurrence matrices. The features obtained after the first four convolutional layers of the reduced variant of pre-trained VGG-16 were used. The improvement in the evaluation parameters improved only marginally, and were again short of the performance obtained just using the visual features.

The failure of the above two methods indicates that the co-occurrence matrices could not model the underlying semantics appropriately. Hence, the next experiment was to ensemble the models instead of fusing the features in the form of a co-occurrence matrix. For the visual features, the earlier LDA model was retained, while for the textual features, an SVM classifier was trained to predict the class based on textual features alone. The classification accuracy obtained by the SVM was about 66%, and this was used to boost the performance during image retrieval.

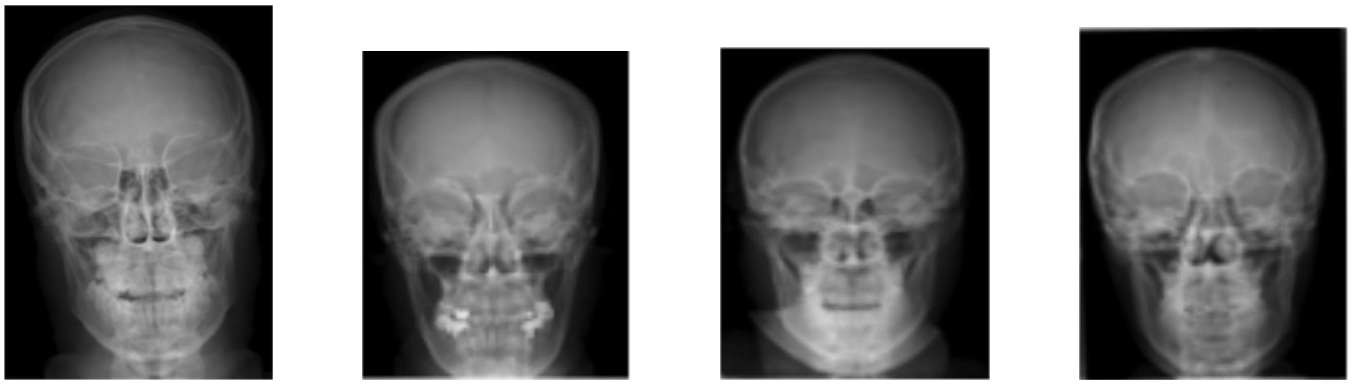
In the proposed ensemble model, when a query image is obtained, firstly, the visual words of the image are computed and compared with the visual features of the images in the dataset, to obtain a ranking of the images. Then, the class of the query is predicted by the SVM using the textual features alone. The ranking obtained using the visual features are re-ranked according to the class predicted by the SVM, the images in the dataset belonging to the same class as the one predicted by the SVM were pushed up while the ones belonging to different class were pushed down. This naturally improved the performance when compared to just the visual features (which had in turn done better than all the co-occurrence based experiments) as we are using additional textual information classifier to reorder the rank list. We have extensively evaluated the retrieval performance using several standard IR metrics, which are described next.

4.1 Retrieval Performance Evaluation

A sample query and the corresponding top 3 documents retrieved are shown in Figure 4. The various metrics used for retrieval evaluation are mAP (Mean Average Precision), gmAP (Geometric Mean Average Precision), Precision@k and NDCG (Normalized Discounted Cumulative Gain). All these metrics are computed as the average obtained over 100 random queries. The results of these experiments are tabulated in Table I.

mAP: The average precision was calculated by using the area under curve of the precision-recall curve (using the 11 point scale), averaged over 100 random queries. The mAP obtained using just the visual features was 0.283, this reduced to 0.155 when the co-occurrence matrix with textual features was used. The autoencoder features and VGG features marginally improved the mAP to 0.16 and 0.162 respectively, which was still off the mAP obtained by just the visual features by a large margin. Finally, from the ensemble approach experiments, we observed that the mAP performance improved significantly, to 0.326.

gmAP: The average precision was calculated in the same manner as that of mAP. The geometric mean was taken over 100 random queries, instead of the arithmetic mean as described earlier. It was observed that the gmAP obtained with just the visual features (0.226), and is more than that obtained in the ensemble method (0.159), even though the ensemble



(a) TEXT: skull query **(b)** First image retrieved **(c)** Second image retrieved **(d)** Third image retrieved

Figure 4. A sample query and top 3 documents retrieved

method has a higher mAP. This shows that the ensemble method has a higher variance when compared to the one model using just visual features. Hence, the ensemble model performed much better when the text classifier predicted the correct class, while doing equally bad when the classifier predicted the wrong class. Even with these limitations, the ensemble method still outperformed the co-occurrence based approaches, which was as expected.

Precision@k: Precision@k is a metric that is used to judge the relevance of the top few documents returned by the model as highly similar to the query image. We performed experiments to evaluate the $top - k$ retrieval performance at k= 5, 10 and 20, using the precision@5 (p@5), precision@10 (p@10) and precision@20 (p@20) metrics. Surprisingly, the approach with just visual features outperformed the ensemble approach by a small margin. Further, both the Visual-features-only model and the ensemble approach outperformed the co-occurrence based approach.

NDCG: NDCG gives a higher score for a ranking list with relevant documents at the top and the non-relevant documents at the bottom. The approach using just the visual features and the ensemble approach again outperformed all the co-occurrence based approaches. The ensemble approach's NDCG was 0.722 and it marginally outperformed the Visual-features-only approach whose NDCG value was 0.702. These results are summarized in tables 1 and 2 as well as in Fig. 5 and Fig. 6.

We compare our work with the state-of-the-art model proposed by Cao et al [2] who also used the ImageCLEF 2009 dataset. Our proposed approach outperformed [2] on both fronts - just visual model and multimodal feature model. Cao et al reported that their visual approach obtained an mAP of **0.0101** which is a lot lesser when compared to the mAP of **0.283** achieved by our model. This clearly indicates that our LDA based vocabulary modeling was much more effective in representing the visual features than PLSA which is used by

Table 1. Comparison of 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**

Table 2. Comparison of performance of various models using precision@5 and 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

Cao et al [2]. Our proposed Late fusion multimodal approach obtained an mAP of **0.326** which outperformed Cao et al's multimodal approach, with its mAP of **0.2909**. There could be a small margin of error in this comparison as the query sets used may be different, but the large difference in mAP especially in the visual performance shows that our proposed approach models the visual features more effectively. This comparison has been illustrated in Fig. 7.

5 Conclusion and Future Work

In this paper, a multi-modal medical image retrieval approach that incorporates both visual and textual features for improved image retrieval performance is presented. In the discussed model, SIFT features are used for capturing the important visual features of the medical images and

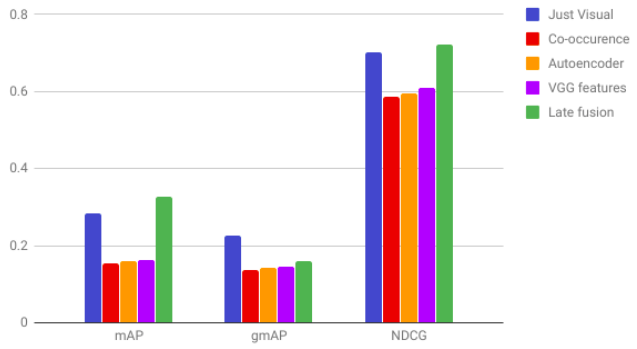


Figure 5. Comparison of mAP, gmAP and NDCG

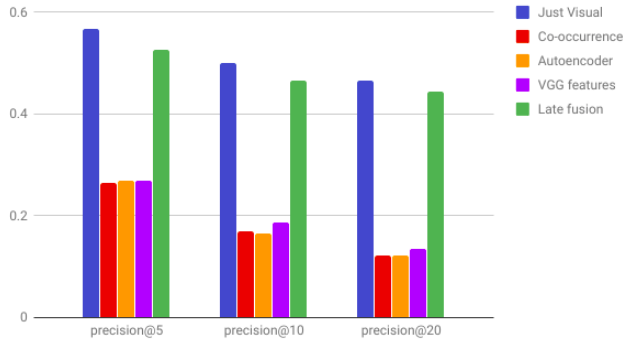


Figure 6. Comparison of precision@5 and precision@10, precision@20

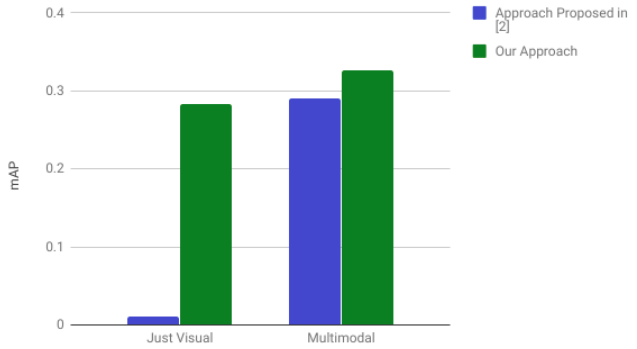


Figure 7. Comparison of proposed model's mAP performance with Cao et al's approach [2]

Latent Dirichlet Allocation (LDA) is used to effectively represent the topics of the clustered SIFT features. To derive the composite feature set, two different fusion techniques were experimented with - early and late fusion. In early fusion, features obtained from an autoencoder and a modified VGG-16 model were used. The late fusion approach was implemented as an ensemble of both visual and textual features, aided by a SVM based classification for improving retrieval performance. Experiments showed that the drop in performance when the textual features are incorporated indicates that the co-occurrence matrix was not an effective way of fusing the textual and visual features in this case. Further attempts to decrease sparsity using autoencoder and using VGG features did not improve the performance. Separating out the textual and visual components using the late fusion approach gave better results. The performance with visual-features-only model was improved by re-ranking the result list using an independently trained text classifier. This outperformed the early fusion approaches proposed in this work as well as those described in other contemporary works such as [2]. 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. We are also working on extending our late fusion approach so that it can be applied to larger corpora.

6 Acknowledgement

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions. This work is supported by the Department of Science & technology, Science and Engineering Research Board, India under Grant No.: 61273304.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[2] Yu Cao, Shawn Steffey, Jianbiao He, Degui Xiao, Cui Tao, Ping Chen, and Henning Müller. 2014. Medical image retrieval: a multimodal approach. *Cancer informatics* 13 (2014), CIN–S14053.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 248–255.

[4] Hayit Greenspan and Adi T Pinhas. 2007. Medical image categorization and retrieval for PACS using the GMM-KL framework. *IEEE Transactions on Information Technology in Biomedicine* 11, 2 (2007).

[5] Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 289–296.

[6] Yonggang Huang, Jun Zhang, Yongwang Zhao, and Dianfu Ma. 2010. Medical image retrieval with query-dependent feature fusion based on one-class SVM. In *Computational Science and Engineering (CSE), 2010 IEEE 13th International Conference on*. IEEE, 176–183.

[7] Thomas Lehmann et al. 2003. The IRMA project: A state of the art report on content-based image retrieval in medical applications. In *Korea-Germany Workshop on Advanced Medical Image*. 161–171.

[8] Rainer Lienhart, Stefan Romberg, and Eva Hörster. 2009. Multilayer pLSA for multimodal image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 9.

[9] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[10] Sandy A Napel, Christopher F Beaulieu, Cesar Rodriguez, Jingyu Cui, Jiajing Xu, Ankit Gupta, Daniel Korenblum, Hayit Greenspan, Yongjun Ma, and Daniel L Rubin. 2010. Automated retrieval of CT images of liver lesions on the basis of image similarity: method and preliminary results. *Radiology* 256, 1 (2010), 243–252.

[11] Trong-Ton Pham, Nicolas Eric Maillot, Joo-Hwee Lim, and J Chevallet. 2007. Latent semantic fusion model for image retrieval and annotation. In *Proceedings of the 16th ACM conference on Conference on information and knowledge management*. ACM, 439–444.

[12] Joel Pyykkö and Dorota Głowacka. 2016. Interactive content-based image retrieval with deep neural networks. In *International Workshop on Symbiotic Interaction*. Springer, 77–88.

[13] Adnan Qayyum, Syed Muhammad Anwar, Muhammad Awais, and Muhammad Majid. 2017. Medical image retrieval using deep convolutional neural network. *Neurocomputing* 266 (2017), 8–20.

[14] Md Mahmudur Rahman, Sameer K Antani, Rodney L Long, Dina Demner-Fushman, and George R Thoma. 2009. Multi-modal query expansion based on local analysis for medical image retrieval. In *MICCAI International Workshop on Medical Content-Based Retrieval for Clinical Decision Support*. Springer, 110–119.

[15] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[16] Xiaoying Tai and Weihua Song. 2007. An improved approach based on FCM using feature fusion for medical image retrieval. In *Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on*, Vol. 2. IEEE, 336–342.