Chapter 2

Literature Review

In the current era of ever-growing web data, single modality based retrieval such as text-based or visual feature-based retrieval is not capable enough for an optimal search result. To search through such enormous data which are of various types, multimodal retrieval has proven to be an effective solution. In other words, multimodal retrieval is a new paradigm for image search. A vast amount of work can be found in the literature for dealing with various kind of modalities for retrieval. Annotation-based image retrieval (ABIR) simply uses text retrieval techniques on textual annotations of images whereas Content-based image retrieval (CBIR) retrieves images by image contents only. Kilinc *et al.*[48] use image annotations to retrieve images from web pages. In their work, image annotations are modified and enriched by surrounding textual content available. A re-ranking approach is also proposed to improve retrieval performance, but image contents are not considered

in this work. Such text-based-only or annotation-based image retrieval methods suffer when annotations are missing. In the real world, it is hard to expect all images being uniformly annotated. For some, annotations may be missing while others may contain noise. In such cases, content-based image retrieval system comes to the rescue. Yoo *et al.* [102] propose a content-based expert system using low-level features for image retrieval but they completely disregard any associated text. Yildizer *et al.* [100] propose a fast and efficient CBIR system. Daubechies' wavelets transformation is used to extract feature vectors from the images. Multi-class Support Vector Regression model is applied on those extracted feature vectors for dimension reduction. Finally, the low dimensional feature vectors are classified by a Support Vector Machine (SVM) based classifier which categorizes the entire image databases into different classes. When any query image comes into the system, the categorized image space reduces the searching time and boosts the searching efficiency. The CBIR systems suffer from the drawback that they do not leverage the benefits of associated text. Current research trend has gradually shifted towards multimodal retrieval which articulates the fact that both text and image features facilitate enhanced retrieval performance.

Buffoni *et al.* [11], present a learning to rank framework for text-based image retrieval. In this learning-based ranking framework, a score function is first learned to extract textual and visual similarity. Three types of similarity scores– scalar product, cosine similarity, and histogram distance are computed to calculate the similarity between descriptor vector and the document descriptor vector. Then the

similarity scores are combined to produce a final ranked list. Wang *et al.* [92] propose a probabilistic model based robust framework for crossmodal and multimodal retrieval. This retrieval framework is capable of combining any media. For combination strategy, the system uses a first-order Markov chain to find the similarity between low-level content and high-level semantics. Thus, the heterogeneous similarity measures are also achieved for different unimedia types. Caicedo *et al.* [12] state that finding a relationship between image content and accompanying text description is not a trivial task. Here, the authors have suggested a Latent Semantic Kernel-based approach to correlate free text and visual features that allows modeling complex document representations by operating with appropriate similarity measures. Lienhart *et al.* [55] capture the concept of combining different modalities through a layered or hierarchical topic-based description of image compositions. They propose a multilayer pLSA model that can handle different modalities at each level. The pLSA model takes separate sets of data from different modalities and keeps them in different leaf level nodes (topics) and then merges the knowledge acquired from these leaf level topics to form a higher level node (topic). Apart from overcoming the challenge of multimodal combination, they show that text+image retrieval outperforms unimodal retrieval significantly.

Out of these mentioned works, we are particularly concerned with the ones combining textual and visual modes. For example, in the paper by Song *et al.* [84], the authors address the problem of combining multiple features to enhance the multimodal retrieval ability. The authors present two fusion strategies, multi-feature

fusion and multi-similarity fusion, where multi-feature fusion, similar to early fusion, concatenates the features after proper normalization and multi-similarity fusion, like late fusion, combines multiple unimodal similarities derived from different distance measures. The multi-similarity fusion, over multi-feature fusion, alleviates the curse of dimensionality of feature concatenation. Experimental results of this work reinstate that fusion strategies indeed enhance retrieval performance. Myoupo *et al.* [65] combine query reformulation and visual image re-ranking for multimodal retrieval. Here, the authors agree upon the fact that textual queries are essential to multimodal image retrieval. Accordingly, concepts semantically similar to the original query are extracted from Wikipedia by neighborhood method and are used for reformulating the query. Also, the authors re-rank the retrieved images on the basis of visual features. Similarly, Moulin *et al.* [64] argue that textual and visual features cannot have the same weight in linear combination. The authors established the fact that a text descriptor is usually better than a visual one. In light of this, the authors adapt two methods– MAP and Fisher Linear Discriminant Analysis (Fisher-LDA) to learn the optimal combination parameters. Experimental results establish that Fisher-LDA approach outperforms other approaches.

An efficient indexing facilitates fast and accurate information retrieval. For text-based image retrieval, linguistic indexing *i.e* image annotation plays a key role in the retrieval task. To aid the detection or annotation of millions of images available online, a number of machine learning or statistical learning based techniques exist. However, learning accurately by these techniques, a decent volume of labeled data

or manually annotated data is required and as stated earlier it is rather unwise to annotate a significant amount of data manually. To overcome this problem, Russell *et al.* [76] develop a Web-based image annotation tool for instant web image annotation. The authors build a large collection of images with ground truth labels to aid the supervised learning and quantitative evaluation. After collecting a large number of annotations for widely spanned object categories, the quality of the image dataset is enhanced through heuristics to recover object parts and depth ordering by employing WordNet.

Image retrieval in response to textual queries requires some knowledge of the semantics of the image [31]. Feng *et al.* [31] show that automatic image annotation and image retrieval through text query can by done together by using a multiple Bernoulli relevance model. The model learns from a given labeled training set of images or even videos where manually labeled corresponding keywords are provided.

The relevance model is a joint probability distribution of the word annotations and the image feature vectors and is computed using the training set. The word probabilities are estimated using a multiple Bernoulli model and the image feature probabilities using a non-parametric kernel density estimate.

Although annotations have been treated as an effective tool for image retrieval, often these human-provided keywords are far from a comprehensive description of the image content and this significantly limits the efficiency of image search. So, Yang *et al.* [99] propose a method to supplement semantic image descriptions by associating

a group of property tags with each existing tag. The authors incorporate six different properties (location, color, shape, texture, size, and dominance) of the detected image objects to refine the existing tags based on their context and establish that those expanded tags with supplementary description enhance the retrieval results.

In the paper by Nguyen *et al.* [66], image annotation problem is studied in a multimodal framework where both visual and textual information are available. The authors propose a Latent Dirichlet Allocation (LDA) based approach where the visual data, the tag data and the topic data are labeled separately and then are associated with the image. LDA is use here to maintain the consistency between the topic generated from the visual data and the topic generated from the text *i.e.* tag data.

If the training set consists of both annotations from experts and also noise in the form of user-generated tags from social media it becomes difficult to ascertain the correct annotation for an image. To alleviate this problem Uricchio *et al.* [87] propose a robust label propagation framework based on Kernel Canonical Correlation Analysis (KCCA), which builds a latent semantic space where correlation of visual and textual features are well preserved into a semantic embedding. Li *et al.* [53] argue that linear regression is an ineffective strategy for automatic image annotation because of its computational complexity. Hence, the authors propose a new automatic image annotation method based on data grouping. Here the images are grouped on the basis of their learning from expert systems employing softmax gate

network of training samples. Zhang *et al.* [105] put forward a new framework for automatic image annotation of regions through segmentation based semantic analysis and discriminative classification. Due to the problem of "semantic gap", generative methods for image annotation often suffer from the images with similar visual features but different semantics. Discriminant models have been used to overcome such issues. Ji *et al.* [46] thus propose a novel image annotation approach which combines the generative and discriminative models through local discriminant topics in the neighborhood of the unlabeled image.

For any task involving multiple features both feature selection and feature combination (considered to be part of feature engineering) are equally important[1][2]. Combining features have been successfully applied to various domains of image and text including text classification [69], text retrieval [1], image classification [3], optical character recognition [18] among others. Image classification using feature combination has been in vogue for quite some time now. Gehler and Nowozin [33] have studied several models aimed at learning the correct weighting of different features from training data. Additionally, they have derived ensemble methods for multiclass object classification.

Plenty of work has been done on image classification but many of them consider only single modality. Diamant *et al.* [26] propose a bag-of-visual-word method for image classification. A mutual information based approach is used to select the most

[1]https://conferences.oreilly.com/strata/strata2012/public/schedule/detail/22489
[2]https://blog.bigml.com/2014/12/02/the-importance-of-feature-engineering/

significant visual-words for better classification performance. The work by Ojala *et al.* [68] focuses on the texture measures as a novel approach for the classification task. In the paper by González *et al.* [35], the authors use multi-resolution pattern instead of local or single resolution pattern for image analysis, even though the classification task remains unimodal.

As far as content-based image retrieval (CBIR) is concerned, Deselaers *et al.* [25] have carried out an extensive study on feature selection and combination strategies. It is noteworthy that most of the strategies mentioned in the papers above are concerned with image features only. However, as Myoupo *et al.* [65] and Datta *et al.* [20] have shown, embedding textual features enhances the performance of text-based image retrieval. Similarly, in all aspects of image related tasks, it has been established that multimodal approach performs better than unimodal one [38][44][67]. The very essence of multimodal approach relies on how well the identified features are combined. Consequently, we take up this task in light of image classification task employing an artificial intelligence approach.

2.1 Literature Gap

While surveying the literature, we observe that there are mainly three focal research points of Multimodal Information Retrieval. A brief discussion of them follows based on our observation.

- The first problem is to bridge the semantic gap between the high-level information need of users and commonly employed low-level features. As we have discussed earlier, it is very hard to express the exact information needs in machine intelligible language for any naive user. It is also found from a study that generally users type only a few words to express their needs and usually the average query length confines within 3 to 4 words long [8]. So it becomes a real challenge to capture the exact information needs from these limited sources of information. Also when the search process includes some multimodal documents *i.e.* other than text like images or videos, the complexity arises much. For image search, it is hardly possible to express an image through words. On the other hand, search through image content is also not always wise. We run the risk of missing out the sense conveyed by the image. Thus the semantic gap between intra-modal or inter-modal sources of information as well as the gap been the users and employed search features are still open problems.

- After selecting the modalities, the next challenge is to correctly represent the varied natured data to aid multimodal search. Many works have been done to come up with a joint representation for different modalities or features. However, there still exists a lot of scope to improve the efficiency or performance. Apart from common representation there comes another problem while dealing with complex features, like the dimensionality problem. Most of these complex feature representations are of higher dimension and they suffer from "Curse of

dimensionality". Not only these higher dimensional features raise the computational complexity but also it makes the process slow. So to overcome this problem, there still needs a vast amount of research to be done.

- The next one is how to determine the optimum fusion strategy. While working with different modalities of data the first issue is to choose the appropriate features from the data of various nature which may aid the search task. After determining the suitable feature, the next challenge is how to combine them as they are of different natures. Also, the weight assignment or the combination strategy plays a vital role in the retrieval performance. Thus, (a) What to fuse, (b) How to fuse and (c) When to fuse (Early or Late fusion) are the three big questions that concern all these multimodal tasks.