Recent years have witnessed the gigantic growth of multimedia contents on the web. Textual data, description of the multimedia document either in the form of metadata, keywords, annotation, associated text etc. transform the web into a truly multimodal in nature. This plethora of multimodal information requires sophisticated models for their representation, indexing and retrieval. However, the presence of the giant semantic gap between human understandable semantics and the machine understanding becomes the bottleneck issue for both the researchers of industry and academia. Most of the researchers rely on multimodal fusion technique i.e. combination of information from multiple modalities to cope with the semantic gap issue. In this direction, three fusion architectures have been adopted in recent years i.e. Intra-modal fusion (decision fusion), Inter-modal fusion (early fusion) and trans-media fusion (Intermediate fusion). Every fusion architecture has their own pros and cons. Intra-modal fusion better cope with the higher-level semantic decisions as they combine the decisions made from multiple classifiers, and provides simple as well as scalable solutions. However, this fusion architecture lacks in handling abstract or derived concepts and also restrict user to provide query in the same combination of modalities as available in the repository, which is not always the suitable case for the user. On the other hand, Inter-modal fusion architecture provides better abstraction capability and flexibility to user in order to provide query in any modality. However, they highly lacks in understanding the user’s dynamic need. Consequently, the composition of former fusion techniques seems to be a
pleasant idea to enhance fusion engine, thus hybrid fusion comes into existence, in order to take the advantages from both the fusion architecture in a cost of putting restriction on user to provide multimodal queries.

We start by proposing a novel query adaptive fusion model namely Inter-multimodal fusion that provides better abstraction capability as well as flexibility to user in essence of accepting whatever the modality or combination of modalities user have as a query. Additionally, the proposed fusion model provides capability to better cope with the multi-level semantics, thus offers broader scope of implementation. Extensive experimental evaluations show that our approach achieves significant improvement over state-of-the-art fusion approaches on two standard and challenging datasets for textual query task. We follow the proposed inter-multimodal fusion for a deeper understanding on handling imperfections in semantic analysis, as it highly influences the discriminative power of semantic decisions.

To cope with imperfections at the query level, we propose query transformation engine namely modality adaptive query transformation (MAQT), which refines user need by transforming uni-modal query into a multimodal query where modality ensures maximum discrimination capability. Extensive experimental evaluations show that our approach achieves state-of-the-art results on several standard and challenging datasets.

It is well known that one of the semantic characteristics of a concept is its contains synonymous representations. While, each document must possess only one representation, hence the presence of imprecision in semantic decisions is obvious, to take care of imprecisions at decision level, a generic framework Bokhari-IMGR: a graph-based inter-multimodal fusion approach is proposed, which is based on the weighted opinion of inter-multimodal fusion. Extensive experimental evaluations show that our approach achieves significant gain over state-of-the-art models on several standard and challenging datasets for both textual and image queries.
Finally, we conclude the work presented in this thesis with discussion on the main achievements of our work followed by its limitations and then points out some perspective directions for future research.