Chapter 2

LITERATURE REVIEW

2.1 BACKGROUND

Recent proliferation of multimedia contents, various multimedia repositories, ubiquitous computing devices and inexpensive availability of storage media to accumulate vast amount of multimedia contents, have necessitated the surge of intelligent maneuverings for smart access to this multimedia data. The branch of Computer Science that facilitates smart access to large collections of multimedia data is called Multimedia Information Retrieval (MIR). It is the process by which a similar collection of multimedia data (image, audio, video etc.) is represented, matched and retrieved for the purpose of fulfilling the need of the user (query). To accomplish this need, earlier approaches of multimedia retrieval rely on textual keywords to represent multimedia documents, these textual keywords later matched with the textual keywords of user’s query. This process of multimedia retrieval is known as Text-based Information Retrieval (TBIR). The main limitation of TBIR is that textual keywords are limited in quantities, which unable to represents all the semantic cues of multimedia contents. To overcome this issue, efforts have been taken towards Content-based Multimedia Information Retrieval (CBMIR), where multimedia contents are represented and stored in low-level feature representation called mode of representation (e.g. images could be represented using texture features, color features, shape features etc.). In online phase, user’s query would also represented in similar low-level feature representation, which then matched with the respective feature representations of multimedia documents. For several decades, the multimedia community has been interested in designing multimedia search systems in order to access multimedia collections. Nevertheless, a core challenge that still makes multimedia
information retrieval an open issue is the “semantic gap”. On the one hand, multimedia data such as images, audio, videos, are stored in machines into a computational representation, which consists of low-level features. On the other hand, humans who search digital collections perceive multimedia contents using high-level concepts such as keywords. It is a difficult task to map both representations in order to match the information need of a user and the items of the collection. To tackle this issue, experimental studies conducted in last two decades shows that decision-making based on the combination of multiple modalities increases the accuracy of the overall decision-making process [Atrey et al., 2010], which would otherwise not be possible by using a uni-modal data. Hence, multimodal information retrieval has gained ample consideration of various researchers for numerous multimedia analysis tasks.

2.2 Multimodal Information Retrieval

Multimodal data from large digital libraries to the web content, transform the digital world with full of multimodal information both in the context of our professional or personal activities. Advent of ubiquitous computing environment, web 2.0 and smart devices have led to provide an ease in both consumption and production of huge amount of multimodal data in our day to day life. This plethora of multimodal data requires efficient methods for their representations and retrieval. The process by which a collection of multimodal data (image-text, audio-visual, audio-video etc.) is represented, matched and retrieved for the purpose of fulfilling the need of the user is called Multimodal Information Retrieval (MMIR). This process involves various stages initiated with representing data and ending with returning relevant information to the user. The complexity in this process is due to the heterogeneous nature of multimodal data representations, due to which, there does not exist any correspondence between their representations and thus there is no way to directly measure the contents of the two modalities. Therefore, a process is required that takes the responsibility to integrate multiple media, their associated features, or their intermediate decisions into a unified representation, that essential sub-process of MMIR is referred as multimodal fusion [Atrey et al., 2010]. The main objective of fusion process is to obtain valuable insights
about the data, a situation, or a higher-level activity extracted from multiple sources of data. To accomplish this objective, a restriction on user is asserted to provide query in the same combination of modalities as available in the search repository. Bounded with this assertion, researchers utilize multimodal fusion at two levels:

1. Feature Level or Early Fusion or Query Independent Fusion
2. Decision Level or Late Fusion or Query Dependent Fusion

In feature level fusion, features extracted from multiple modalities are first concatenated into a unified feature vector and then the analysis task would perform on combined feature vector. In other words, multimodal features extracted from repository as well as from users query are first combined into respective unique representation and then the matching between their contents would take place. The advantages of feature level fusion are of two folds: 1) it can utilize the correlation between multiple features from multiple modalities at an early stage that helps in better task accomplishment. 2) It requires a single classifier or model to perform an analysis task on the combined feature vector [Snoek et al., 2005]. Apart from the pros side, feature level fusion has cons side also. Two highly influential issues with feature fusion are: 1) the increase in the number of modalities makes it difficult to learn the correlation among the heterogeneous features. 2) Increase in the number of features, leads to curse of dimensionality and thus becomes unrealistic for implementation purposes. To overcome these challenges, advances have been reported in multimodal retrieval systems [Paramita et al., 2010; Tsikrika and Kludas, 2008; Smeaton et al., 2006; Snoek et al., 2005b; Datta et al., 2008; Iria et al., 2009]. These are extension of previous systems, where a common retrieval system combines information from multiple modalities and then apply feature reduction techniques on the concatenated feature vectors. One popular approach is to concatenate features from different modalities and rely on unsupervised latent factor analysis like latent semantic analysis (LSA). In general, the inability to access each data modality individually (after the fusion of modalities) limits the capability of these systems for generic retrieval scenario.

On the other hand, in late fusion or decision level fusion, multiple classifiers or models were used to predict their local decisions, later the fusion engine generates a final
decision on the basis of these local decisions. Decision level fusion is one of the most common and widely accepted forms of fusion due to its simplicity and performance. As, the information received by the fusion engines has already been processed there is no need to worry about the noisy data. This means that fusion has to rely on preprocessed information in order to construct semantic meaning from combining partial information coming from individual classifier. The preprocessed information constitutes a hard decision that was produced by one or more classifiers. Those decisions can then be combined using different methods (e.g. majority voting, rank fusion, AND fusion, OR fusion etc.). A good overview of these methods is given in [Escalante et al., 2008]. Nevertheless, the results produced by these methods might be suboptimal due to two main reasons: 1) in general, individual modality contains noisy data and 2) requires a lot of learning efforts. As, every modality needs a separate model for an analysis task.

Durrant-Whyte classified three different strategies for different fusion goals referred as complementary fusion, cooperative fusion and competitive fusion strategies [Durrant-Whyte, 1988]. Complementary fusion exploits the independence among the information sources whereas cooperative fusion strategy exploits the dependence among the information sources. While competitive fusion strategy is much related to expert selection. These fusion strategies influence the way fusion process takes place under diverse application settings. However, in real application scenario, it is not known in priori that which fusion strategy provides better performance results to fulfill the user’s semantic need. To satisfy the user’s need, the techniques of adaptive multimodal information retrieval are used.

2.3 Adaptive Multimodal Information Retrieval

Adaptation is an important research direction for information retrieval in general and multimodal systems in particular. Adaptive systems have the capability to modify their working strategy in diverse application scenarios by adapting the environment factors. This capability of adaptive systems broadens their scope of implementation and execution in real world applications. Due to these benefits, adaptive systems have been extensively applied in numerous of research domains including multimodal information
retrieval per se. The main objective behind the development of adaptive multimodal information retrieval (AMMIR) systems is to provide a platform for erecting embedded systems that capture, reason and act upon user intentions [Tsiporkova et al., 2013]. This process requires adaptation capability in two different components of systems i.e. interface and retrieval. In order to seamlessly integrate the two components, retrieval engine must have to be flexible enough to absorb any possible combination of inputs from interface and restricted enough to deliver multimodal outputs to interface. Later, interface has to adapt upon users environmental settings. Adaptive Interface design and Adaptive information retrieval are two parallel research fields that are extensively studied nowadays. This work is related with adaptive information retrieval in general and adaptive multimodal fusion in particular. Hence, adaptive interfaces are out of the scope of this thesis.

The benefits of adaptive multimodal fusion arrives with a certain cost and complexity in the retrieval process. This is mostly due to the different semantic characteristics of components involved in the retrieval process including queries, documents, class settings, modalities, and domain etc., some of them are briefly stated in the following [Atrey et al., 2010; Clinchant et al., 2011]:

— The heterogeneous nature of multimodal data representations. Different modalities exhibits different level of semantics in their content representation e.g. image modality contain low-level feature representation like SIFT, HOG, SURF, BOVW etc. which contains low-level semantics while textual modality usually contain high level representations that human perceive in his/her day to day life like BOW, LDA etc. containing different dimensions of features in both modality.

— The modalities may be dependent or independent. The dependence between the modalities can be observed at different levels, such as the dependence among low-level features and the dependence among semantic-level decisions. On the contrary, independence among the modalities also provides additional cues in obtaining a decision. When fusing multiple modalities, this dependence and independence information may provide valuable insight based on a particular scenario or context.
— The different modalities generally have varying confidence levels in accomplishing different tasks e.g. textual modality usually have higher confidence in analyzing abstract level concepts like beach, animal etc. while image modality contribute more in concrete or visible concepts detection like frog, banana etc.

— Different classes within a domain generally have varying level of semantic information e.g. a library science domain may contains several books of different subjects etc.

— Different domains generally exhibit varying level of semantic information.

— Different modality query generally have varying semantic confidence e.g. image queries are ambiguous for humans from a semantic viewpoint while textual query are closer to users’ semantics.

The above characteristics of various components influence the way adaptation process is carried out. Due to these varying characteristics and the objective task that need to be carried out, numerous challenges may occur in adaptive multimodal information retrieval:

— What to Fuse?

Earliest consideration of researchers in the field of multimodal fusion is to decide what strategy to follow when fusing multiple modalities. The most widely used strategy is to fuse the information at the feature level, which is also known as early fusion. The other approach is intermediate level fusion or factor fusion, which fuses multiple modalities based on their latent characteristics. Third category is decision level fusion or late fusion [Hall et al., 1997; Snoek et al., 2005], which fuses multiple modalities in the semantic space. A combination of early and late fusion approaches is also practiced as the hybrid fusion approach [Wu et al., 2006].

— How to Fuse?

There are several methods that are used in fusing different modalities. These methods are particularly suitable under different settings. It is important to resolve how the fusion process utilizes the feature, factor and decision level correlation among the modalities.
What to Adapt?

Context can be defined in various ways like situational context, social context etc. which may change from situation to situation or remain stable or sometimes one contextual aspect becomes important for a particular situation while becomes vital for some other situation.

2.3.1 AMMIR Parameters

The adaptive fusion parameters constitute the source for adaptation in multimodal fusion scenarios. Researchers have used various parameters, like level of fusion (e.g., feature, intermediate, decision), fusion strategies, diversity, correlation, scalability, complexity, independence, imperfections, context adaptation etc. [Kokar et al., 2004; Kludas et al., 2008; Kraft et al., 2005; Atrey et al., 2010; Znadia, 2014] for providing information extracted from multimodal data. Most of them are specialized for particular settings. For the sake of generality, only one parameter is considered in this work as:

— Context-of-Retrieval (e.g. representational context, interactional context etc.)

2.3.1.1 Context-of-Retrieval

Humans have inherent capabilities to draw inferences from implicit situational information or understanding of context, which results in practicing good communication. However, this aspect is generally missing in human-computer interactions (HCI). To enhance the utility and effectiveness of communication concerning human-computer interaction, it is suggested to improve the systems’ understanding about context.

The term “context” has been defined in several ways, but from the perspective of practical implementation in HCI, the definitions by [Schilit 1994] and [Pascoe 1998] is important to consider. According to Schilit, information related to where you are, whom you are with, and what resources are nearby may be used as context [Schilit et al., 1994]. According to Pascoe, context may be the subset of physical and conceptual states of interest to a particular entity [Pascoe, 1998]. Hence, the whole situation relevant to an application, environment and its set of users may be considered as the context. Further, the context may change from situation to situation and sometimes one aspect of the
context becomes important and some other aspect may be vital at some other point of time. The terms “situation” and “context” are often used interchangeably; these terms refer to some kind of external or latent factors affecting users’ or system’s behavior. In particular, sometimes context may specify fine descriptors, such as time and noise, while some other time may refer to abstract descriptors like events, locations, surrounding entities, and so forth.

The Vildjiounaite taxonomy recommends five abstract levels of contextual factors for lightweight adaptation, namely, historical factors, social factors, task factors, environmental factors, and computational factors [Vildjiounaite et al., 2015]. Those are explained below:

— **Historical factors** contain past experiences that may affect the current state of system, for example, previous knowledge, user or system past actions, changes in user’s preferences or appearance over time, and recent search logs.

— **Social factors** may comprise of rules and customs of interaction between entities, for example, conversation with an elder etc.

— **Task factors** include users’ intents, for example, purpose of information search.

— **Environmental factors** are anything in surroundings that may affect decision making, for example, background noise and presence of some other object etc.

— **Computational factors** can be anything that provides current system settings, for example, availability or quality of a certain data structure (such as image resolution).

The major research challenges related to context adaptation may be outlined as follows:

- *What are relevant factors of context that influence fusion?*
- *How can these factors be captured?*
- *In what way should content and services be adapted to these factors?*

Adaptation is about to understand the environment and accordingly modify systems action and services to better fit the users intent. The recent upsurge in human-computer interaction has revolutionized the concept of context. Dourish theories of
context classified context as 1) “Context is a representational problem” and 2) “Context is an interaction problem” [Dourish, 2004]. Former theory perceive context as information that can be represented in the form of features like the other information is represented in computer. While, later theory assumes that the context cannot be described with a predefined feature set. Instead, the choice of descriptors is a result of human activity. Hence, contextual factors can be classified into representational contextual factors and interactional contextual factors. Which can be automatically obtained through human-machine interaction. Further, on the same line, adaptive fusion models can be classified as interaction based or query dependent fusion (multimodal fusion) and representation-based or query independent fusion models (inter-modal fusion).

This thesis attempts to answer the research questions, namely. "How can these factors be captured? and "In what way should content and services be adapted to these factors?"

### 2.3.2 AMMIR Approaches

In adaptive multimodal information retrieval, various models are used to adapt the contents and services to the user, based on their intrinsic fusion properties. The following sub-section is to distinguish two major retrieval paradigms in adaptive multimodal information retrieval: multimodal retrieval and inter-modal retrieval.

#### 2.3.2.1 Multi-Modal Retrieval

Most of the techniques used in multimodal fusion can broadly fall into three categories: late fusion, graph-based fusion and transmedia fusion. These methods are described by distinguishing the fundamental concepts they are composed of i.e. linear fusion and non-linear fusion. In multimodal retrieval, this is asserted that the user have to provide multimodal query as similar to any other item of the multimodal collection, that is, a document consists of an image part and a text part. Given a multimodal query, the search process accounts for effectively measure the similarity between the query and the multimodal documents available in the repository.
2.3.2.1.1 Linear Fusion

Linear fusion approaches do not act at the feature level representations but rather at the higher semantic level that finds the similarities between intra-modal similarities [Bruno et al., 2008; Ah-Pine et al., 2008]. Albeit simply, late fusion techniques can be seen as functions that take as inputs the vectors of intra-modal similarities between the query and the documents of the multimodal repository. These technique focuses towards the aggregation of multiple intra-modal similarities either on score level or rank level. In this case, the simplest aggregation technique used is the mean average [Escalante et al., 2008], but more elaborate approaches have been studied in recent years [Müller et al., 2010; Csurka and Clinchant, 2012; Wilkins et al., 2010]. Linear weighted fusion method is the common choice of researchers due to its simplicity and effectiveness. The two variant of linear weighted fusion method are linear weighted sum fusion or linear weighted product fusion. It merges the intra-modal similarities by assigning contextual weights to modalities [Vildjiounaite et al., 2015].

2.3.2.1.2 Non-Linear Fusion

Transmedia fusion methods act like the similarity diffusion processes. Unlike late fusion methods, which implicitly consider documents of the repository independent from each other, these techniques leverage the information conveyed by the similarity relationships between each pair of documents. Therefore, these methods go beyond late fusion approaches by taking into account not only the similarity vectors of the query with the different media components but also the inter-media similarity of the repository documents. Simply, Transmedia fusion techniques mix different intra-modal similarity matrices by means of matrix multiplication operations [Ah-Pine et al., 2015]. In these fusion models, similarity diffusion process usually carry out by using pseudo-relevant items only, which are given by the k nearest neighbors. Furthermore, the transmedia principle seeks to utilize one type of intra-modal similarity to find the similarity for the other one. Thus, these methods can also be understood as a generalization of the intra-modal pseudo relevance feedback mechanism.

Graph-based methods consider multimodal documents as nodes of a graph and the type of relationships that they share is represented as its edges. Examples of weighted
edges between objects are visual similarities or textual similarities, but depending on the application, other types of relations can be considered. Graph analysis techniques are then employed in order to infer new features in the goal of rearranging the text-based ranked list of items. One such method, inspired by the well-known PageRank [Grin and Page, 1998; Langville and Meyer, 2005; Franceschet, 2011], was proposed [Hsu et al., 2007b; Hsu et al., 2007a]. It is based on random walks over a stochastic matrix, which is deduced from the fusion of visual and textual similarities, and the stationary probability distribution over the nodes is then additionally used to re-rank the initial retrieved list. In the same vein, Markov random walk model [Craswell and Szummer, 2007] was proposed with backward and forward steps. They found that the best performances were obtained with a long backward walk with high self-transition probability.

2.3.2.2 Inter-Modal Retrieval

In inter-modal retrieval procedure, users have a flexibility to search various modalities of data including texts, images and videos, by providing any other modality as a query. In this process, cross-modal associations among the inter-modal documents are obtained by learning an abstract representation from various data modalities available in a domain. At last, these common representations enable the cross-modal retrieval to find solutions of search result ranking and summarization. To obtain cross-modal association, there are three broad families of learning approaches, which are as follows: 1) Sub-space Learning, 2) Metric Learning and 3) Manifold Learning. Out of these, sub-space learning and metric learning methods are focused towards maximizing the dependence among multimodal data. While, manifold learning methods are deliberate to minimize the independence among the multimodal data.

2.3.2.2.1 Subspace Learning

The main objective of sub-space learning methods is to maximize the dependence among multimodal documents. The main difficulty of cross-modal retrieval is how to measure the content similarity between different modalities of data. Subspace learning methods are one type of the most popular methods. They aim to learn a common subspace shared by different modalities of data, in which the similarity between different modalities of data can be measured. Unsupervised subspace learning methods use pair-
wise information to learn a common latent subspace across multimodal data. They enforce pair-wise closeness between different modalities of data in the common subspace. Recently, several approaches for establishing inter-modal relationships between data from different modalities generally rely on subspace learning, such as Canonical Correlation Analysis (CCA) [Hardoon et al., 2004; Kim et al., 2007], Partial Least Squares (PLS) [Rosipal and Kramer, 2006] and Bilinear Model (BLM) [Sharma et al., 2012; Tenenbaum and Freeman, 2000]. Specifically, CCA is probably the most popular one due to its widespread use in cross-media retrieval [Rasiwasia et al., 2010; Gong et al., 2014; Pereira et al., 2014], cross-lingual retrieval [Udupa and Khapra, 2010] and some vision problems [Li et al., 2011; Rasiwasia et al., 2010] address the cross-modal retrieval problem by investigating the correlations between two modalities, where CCA is proved to be effective. Li et al. apply CCA to face recognition based on non-corresponding region matching [Li et al., 2011]. They use CCA to learn a common space in which the possibility of whether two non-corresponding face regions belong to the same face can be measured. Recently, Partial Least Squares (PLS) [Rosipal and Kramer, 2006] is also used for the cross-modal matching problem. Sharma and Jacobs [Sharma and Jacobs, 2011] use PLS to linearly map images in different modalities to a common linear subspace in which they are highly correlated. Chen et al. applied PLS to the cross-modal document retrieval [Chen et al., 2012]. They use PLS to switch the image features into the text space, and then learn a semantic space for the measure of similarity between two different modalities. A bilinear model (BLM) [Tenenbaum and Freeman, 2000] to derive a common space for cross-modal face recognition, and BLM is also used for text-image retrieval in [Sharma et al., 2012].

2.3.2.3 Metric Learning

Another perspective for cross-modal retrieval is to learn a metric between different modalities of data. [Li et al., 2003] introduce a cross-modal factor analysis (CFA) approach to evaluate the association between two modalities. The CFA method adopts a criterion of minimizing the Frobenius norm between pairwise data in the transformed domain. [Wu et al., 2010] study the metric learning problem to find the similarity function over two different spaces. [Mignon and Jurie, 2012] propose a metric
learning approach for cross-modal matching, which considers both positive and negative constraints. [Quadrianto and Lampert, 2011] propose a new metric learning scheme to learn projections from the data in different modalities into a shared feature space, in which the Euclidean distance provides a meaningful intra-modality and inter-modality similarity. [Zhai et al., 2013] propose a regularized metric learning algorithm to learn a heterogeneous metric for cross media retrieval. [Lu et al., 2013] and [Wu et al., 2013] study the cross-modal retrieval as a problem of learning to rank. They attempt to learn a metric such that the ranking of the data induced by the distance from a query can be optimized against various ranking measures. Learnt Hamming metric is used to speed up the cross-modal search [Bronstein et al., 2010; Zhen and Yeung, 2012; Masci et al., 2014], but the Hamming metric is discrete-valued so that its retrieval accuracy is lower. These methods mentioned above generally treat similar pairs and dissimilar pairs or rank lists equally when modeling the structure of the multimodal data. However, some less informative pairs and rank lists may potentially lead the model to depart from the correct structure, which degrades the performance.

2.3.2.4 Manifold Learning

[Mahadevan et al., 2011] propose maximum covariance unfolding (MCU), a manifold learning algorithm for simultaneous dimensionality reduction of data from different modalities. [Shi et al., 2012] propose a principle of collective component analysis (CoCA), to handle dimensionality reduction on a heterogeneous feature space. [Zhu et al., 2014] propose a greedy dictionary construction method for the cross-modal retrieval problem. The compactness and modality-adaptivity are preserved by including reconstruction error terms and a Maximum Mean Discrepancy (MMD) measurement for both modalities in the objective function. [Wang et al., 2013] propose to learn the sparse projection matrices that map the image-text pairs in Wikipedia into a latent space for cross-modal retrieval. To overcome these difficulties, progress has been made towards cross-modal systems. This includes retrieval methods for corpora of images and text [Denoyer et al., 2004; Qi et al., 2011], images and audio [Li et al., 2003; Zhang et al., 2007], text and audio [Slaney, 2002], images, text, and audio [Yang et al., 2009; Zhang et al. 2007; Zhuang et al., 2008; Zhuang et al., 2007; Yang et al., 2008], or even other
sources of data like EEG and fMRI [Mahadevan et al., 2011]. One popular approach is to rely on manifold learning techniques [Mahadevan et al., 2011; Yang et al., 2009; Zhang et al. 2007; Zhuang et al., 2008; Zhuang et al., 2007; Yang et al., 2008]. These methods learn a manifold from a matrix of distances between multi-modal objects. The multi-modal distances are formulated as a function of the distances between individual modalities, which allows singling out particular modalities or ignoring missing ones. Retrieval then consists of finding the nearest document, on the manifold, to a multimedia query (which can be composed of any subset of modalities). The main limitation of these methods is the lack of out-of-sample generalization. Since there is no computationally efficient way to project the query into the manifold, queries are restricted to the training set used to learn the latter. Hence, all unseen queries must be mapped to there nearest neighbors in this training set, defeating the purpose of manifold learning.

2.4 Summary

This chapter is intended to present state-of-the-art discussion on multimodal information retrieval followed by adaptive multimodal fusion. In AMMIR, the analyses of contextual parameters used to adapt the fusion process is of paramount importance. The discussion is based on the two parallel theories of context, one claim that context is a matter of representation that can be represented at design time. While, the other theory claim that context can only be identified through user interaction. In this work, we analyze that both the theories of context are complimentary in nature. This attempt leads us to design a self-adaptive multimodal fusion engine that better understand the contextual environment. The discussion also incorporates the theories supporting contextual as an adaptation problem based on these parameters.

During review, our main focus has been on adaptive information retrieval methods and techniques used in literature. This chapter also relates the review with the research challenges discussed in Chapter-1 and presented an overview of the recent works in this area. Though excellent works have been done to overcome the challenges in the area of adaptive multimodal fusion and adaptive retrieval, all of these are inline of either of the one contextual theory, due to which the pioneer work also limited to
experimental lab systems which are not in common practice of universal adaptive system. Further, there appears a clear lack in consensus among researchers to what parameters be used to deliver adaptive multimodal contents. The aim has been to find out how our work can contribute to the solution of this problem, especially to understand and adapt the contextual environment throughout the fusion process in an unsupervised fashion. We have proposed the use of these components as an intermediary source for contextual parameters. This enables us to work on semantic multimodal information retrieval.

Chapter-3 deals with the novel linear fusion architecture based on hybrid fusion approach that provides relevance objects (the most relevant documents) matching the users’ search need. Chapter-4 discusses the proposed query transformation approach that transforms uni-modal query into multimodal query. Chapter-5 discusses the proposed inter-multimodal graph based fusion that enhances overall retrieval performance.