Contextual Query Perfection by Affective Features Based Implicit Contextual Semantic Relevance Feedback in Multimedia Information Retrieval

Karm Veer Singh[1] and Anil K. Tripathi[2]

**[1]Department of Computer Engineering, Indian Institute of Technology, Banaras Hindu University
Varanasi, 221005, India**

**[2]Department of Computer Engineering, Indian Institute of Technology, Banaras Hindu University
Varanasi, 221005, India**

Abstract

Multimedia Information may have multiple semantics depending on context, a temporal interest and user preferences. Hence we are exploiting the plausibility of context associated with semantic concept in retrieving relevance information. We are proposing an Affective Feature Based Implicit Contextual Semantic Relevance Feedback (AICSRF) to investigate whether audio and speech along with visual could determine the current context in which user wants to retrieve the information and to further investigate whether we could employ *Affective Feedback* as an implicit source of evidence in CSRF cycle to increase the system's contextual semantic understanding. We introduce an Emotion Recognition Unit (ERU) that comprises of spatiotemporal Gabor filter to capture spontaneous facial expression and emotional word recognition system that uses phonemes to recognize the spoken emotional words. We propose *Contextual Query Perfection Scheme (CQPS)* to learn, refine the current context that could be used in query perfection in RF cycle to understand the semantic of query on the basis of relevance judgment taken by ERU. Observations suggest that *CQPS* in AICSRF incorporating such *affective features* reduce the search space hence retrieval time and increase the system's contextual semantic understanding.

Keywords: *Contextual semantic relevance feedback, spoken emotional words, affective feedback, facial expression, Multimedia Information Retrieval, video retrieval*

1. Introduction

Bridging the semantic-gap is a core challenge in multimedia information retrieval field [1]. In most scenarios MIR systems are primarily designed with the human being as the user. Semantic and sensory gaps can be tackled possibly well by including the human user in the loop of the system that links signals to symbols. Relevance feedback, for retrieval purposes, utilizes the human's role as a consumer of multimedia information. Based on such feedback information, the system updates a distance metric among the low-level features to better reflect high-level conceptual semantic distances and modify the parametric space, feature space, semantic space, or classification space to identify the relevant and irrelevant information. Various implicit [2,3,4] and explicit [5,6] relevance feedbacks and combination of implicit relevance feedback with explicit relevance feedback [7] methods have been proposed in recent years to deal with information needs of users. Relevance feedback, as an interactive and iterative process, is used to remove ambiguity of user's information needs, but not sufficiently as may be desired. The term semantic concept of some information is the interpretation of low level object captured during the image filtering process into high-level knowledge understandable by the subject. A multimedia semantics are associated with information, sought by a user, and these semantics depend on some context and hence we are exploiting the plausibility of context associated with a semantic concept. Such a semantic concept should be context-driven (or context based). As semantic concept, without context can be added in to itself, has little relevance. To enhance the relevance of semantic concept it should be enriched with the context. For instance, there is a large video repository and among them some videos contain information about cricket match played, some of them contain football match played and some of them contain hockey match played. Suppose somebody is interested in finding a video containing a football match. Suppose all the three types of videos contain stadium with crowd, players in the field and a round object moving in the field. If we segment and classify the various objects by only considering visual similarity, the moving round shape may create ambiguity whether it is football, cricket ball or hockey ball. But, if we consider the context in which user wants to retrieve the information (football) and match first the audio signals of spoken word football, hockey and cricket in all videos, then matching visual similarity, we can find relevant videos containing football match accurately with less retrieval time and search space. Thus exploiting the plausibility of context associated with semantic may be beneficial in retrieving relevance information. A context-driven (or context based) information retrieval will be capable of the adaptation to the current needs and interest of the user, the context of the current task. In [8, 9], authors discuss challenges in a contextual retrieval approach. We, hereby, propose to combine the context and semantic for the purpose of enhancement of the possible relevance and are trying to convey the semantics (in the query) to the system through context description. So, one of the major issue is retrieval of a relevant information and some of major challenges are as following:

1) How to effectively model system's contextual semantic understanding of low level objects from the feedback?
2) How system will identify current context in a query? What mechanism will be used to identify current context in query, memorize or to accumulate the relevance feedback

information to estimate the query point movement under the current context of a query that may be beneficial to improve both the current query accuracy and the future system performance in term of precision, recall as well as reduced search space hence retrieval time.

3) It is suggested that relevance feedback can be utilized for affective retrieval [10]. How, we could employ affective feedback as an implicit source of evidence in RF cycle and how it can be used in Contextual Query Perfection Scheme to increase the system's contextual semantic understanding is another challenge.

Contextual semantics understanding can be improved by firstly either by intrinsic context which have complete implicit knowledge of multimedia and require priori knowledge of objects and availability of good training sample sets or by extrinsic context which have multi-modal characteristics, for example, use of audio and speech to increase the contextual semantics of visual contents. Secondly, contextual semantic understanding can also be improved by extraction of contextual semantics from streams of user actions and emotions on multimedia contents. How the user interacts with multimedia information (eye tracking, emotion, browse and navigation operations)? All these information show the interest of the users on multimedia content, and would certainly be complementary to contextual semantic annotation and multimedia content analysis processes [11, 7]. Multimedia information has multiple semantics, which depend on context, a temporal interest and preferences. To cope with above discussed major challenges, the contextual semantic understanding of low-level objects, we argue that use of audio, speech in determining the context in which user wants to retrieve the information, would increase the system's contextual semantic understanding of visual contents. Audio features in a query clip will be extracted and matched with stored audio features in audio feature database, if found, the visual features in the query clip will be extracted and matched with the corresponding visual features of audio from visual features database. Audio with matched visual features would become the context in which user wants to retrieve the information. Further, we argue that incorporating spoken emotional words (words list is given in Table 1.) along with facial expressions facilitate a more natural and meaningful emotion. Such a combined affective feature would be more effective implicit source of evidence in contextual semantic relevance feedback cycle and could allow a search system to predict, with reasonable accuracy, the contextual semantic relevance of information without the help of explicit knowledge.

we are proposing an *Affective Feature Based Implicit Contextual Semantic Relevance Feedback (AICSRF)* framework to investigate whether audio and speech along with visual could determine the current context in which user wants to retrieve the information and to further investigate whether we could employ *Affective Feedback* as an implicit source of evidence in CSRF cycle to increase the system's contextual semantic understanding.

We are proposing *Contextual Query Perfection Scheme* to learn, refine the current context that could be used in query perfection in RF cycle to understand the semantic of query in following iteration. We investigate whether audio, speech could determine the context in which user wants to retrieve the information and could increase the system's contextual semantic understanding of visual contents. The end goal is to model user's affective responses, understanding the relationship between contextual semantics and emotion (spoken emotional words along with facial expressions). We propose and incorporate spoken emotional words along with facial expressions as an affective feature in feedback cycle and further investigate whether such affective features used as an implicit source of evidence in feedback cycle could increase the contextual semantic relevance.

Table 1: Spoken emotional words

Positive emotion words	Vow, Hurrah, Wah, Yes, Yeah, Oh yes, ok, Beautiful, Superb, Mind blowing, Weldon, Fantastic, Lovely, Lovely scene, Amazing, Excellent, Tremendous
Negative emotion words	No, Oh no, Never again, Oh shit, Nonsense, yaaaah

In section 2, we introduce some recent work related to this research. Section 3 provides the details of our proposed AICSRF framework to investigate whether audio, speech could determine the context in which a user wants to retrieve the information and to further investigate whether affective features can be used as an implicit source of evidence in feedback cycle. We will discuss *Contextual Query Perfection Scheme* which refines the audio-visual contexts on the basis of relevance judgment taken by ERU via relevance feedback. Section 4 reports the experiments conducted and discusses the results. Finally, section 5 concludes this research and points out the future directions.

2. Related Work

To our knowledge, in multimedia information retrieval, increasing the contextual semantic understanding, modeling of affective behavior of user and its use in contextual semantic relevance feedback is a new and less explored area. However, some earlier works exist in several similar fields. The contextual relevance feedback approach has well recognized research challenges in information retrieval. In [12], Author has presented ongoing research on the implementation of the contextual relevance feedback approach in web-based information retrieval. The approach builds a contextual user profile employing the user's implicit data (i.e. from Internet browsing history) and explicit data (i.e. from a lexical database, a shared contextual knowledge base and domain-specific ontology/concepts) to provide relevant information to users that potentially satisfies their information needs. In [13] the relevance feedback is applied to reflect the user's emotion in every retrieval processes. In [14] the authors propose the effective image retrieval method based on the user's emotion. In [15, 7], the authors adopt a user-centered approach and develop models that can infer relevance implicitly from eye-movement data. In [16], the authors explore the role of affective feedback in designing multimedia search system, employing sensory data that derive from facial expressions and other peripheral physiological signals (heart rate signal, galvanic skin response, skin temperature, motion, etc) as the only feedback information. In [17], the authors present a novel video search environment that aggregates information from users' affective behavior, by applying real-time facial expression analysis. Facial expressions have been associated in the past with universally distinguished emotions, such as happiness, sadness, anger, fear, disgust, and surprise [18]. Recent research also indicates that emotions are primarily communicated through facial expressions [19], which provide useful cues (smiles, chuckles, smirks, frowns, etc.) that are considered an essential part of our social interaction [20]. Few efforts have been done to link emotions to content-based indexing and retrieval of multimedia [21]. [21, 22] analyze the text associated to a film searching for occurrences of emotionally meaningful terms; [23] analyze pitch and energy of the speech signal of a film. A set of affective features was extracted from audio content, and was annotated using a set of labels with predetermined affective semantics. In [24], the concept of speech-assisted facial expression analysis and synthesis is proposed, which shows that the speech-driven facial animation technique not only can be used for expression synthesis, it also provides useful information for expression analysis.

3. Overview of the ECSRF framework

We propose, hereby, a method of relevance feedback that makes the system automatically learn contextual semantic concepts using audio-visual context and then retrieves information under a selected context under which a user wants to retrieve the information. The system takes relevance judgments via user's affective feedback (spoken emotional word(s) and facial expressions) in a particular context. Audio and speech present in a query happens to be primary context. More than one context may be present in a query due to different audio, speech present in query clip, hence, semantic concept of visual object varies according to contexts under which user wants to retrieve the information. The system understands contextual relevance by taking affective feedback as an implicit source of evidence. The context is refined iteratively in this contextual relevance feedback process until user is satisfied with retrieved video result sets. The architecture of the proposed AICSRF framework and system workflow is being depicted in Fig. 1. There are mainly four key components namely Audio-Visual Context Generator Recognition Unit (AVCRU), Emotion Recognition Unit (ERU), Contextual Semantic Relevance Feedback Unit (CSRFU) and Contextual Query Perfection Scheme. The typical system output would be semantics of a visual content associated with set of contexts to which they are related and will be stored as contextual semantic database. User gives initial query in the form of a video clip to the system by using QBE. Audio and speech in a query clip are extracted and matched with the trained audio feature sets and speech feature sets DB stored in Contextual-seed Training Database discussed in section 3.1.3. The matched audio or speech becomes the audio context. The visual features in query clip is now matched with the corresponding trained visual feature sets of this matched audio context in Contextual-seed Training Database and becomes visual context. The audio-visual context recognized by the method discussed in detail in section 3.1.4 becomes the initial context under which user wants to retrieve the information from videos. This initial Audio-Visual feature sets formulate the initial query. In retrieval process, first the audio features set of this initial Audio-Visual context is matched with audio and speech features present in video from video repository then the retrieved video result sets are again searched for matching of the visual features set of the initial Audio-Visual context with the visual features present in videos of retrieved video result sets. The final top rank videos are displayed in Browser. User opens some of the videos to see the desired contents. The system continuously monitors the user's emotions on these open contents by applying method discussed in ERU under section 3.3. The ERU discriminates the user's emotions, shown on open contents, into positive and negative emotion and provides the implicit relevance feedback to the system. The system categorizes these retrieved video result sets into positive (relevant) and negative (irrelevant) sample sets based on the positive and negative emotions. The Contextual Query Perfection Scheme takes these positive and negative sample sets and evaluates the contextual fitness of a semantic concept. The Contextual Query Perfection Scheme refines the audio-visual contexts on the basis of relevance judgment taken by ERU via relevance feedback. These refined audio-visual contexts are again used as new contexts to retrieve the video. The whole retrieval process is repeated until user is satisfied with retrieved information. In the following section, we will discuss each individual component in details.

3.1 Audio-Visual Context Generator and Recognition Unit (AVCRU)

Our argument is that the use of the audio and speech along with the visual features in determining the contexts under which user wants to retrieve the information, would increase the contextual semantic understanding of the system about the visual contents present in the videos. The semantic concept of visual content varies according to the context under which user wants to retrieve the information. In this proposed method, audio and speech are becoming the primary context descriptor and visuals are becoming the secondary context descriptor. The combination of these two context descriptors is the combined Audio-Visual context

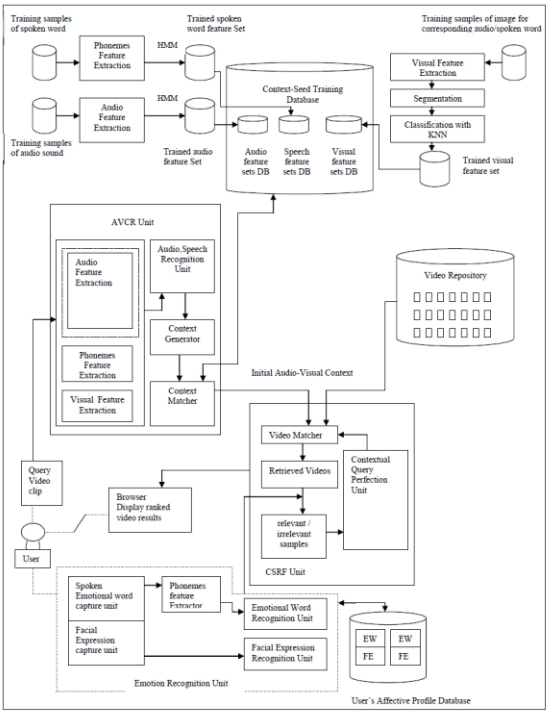
Training samples of spoken word
Trained spoken word feature Set
Training samples of image for corresponding audio/spoken word
Phonemes Feature Extraction
HMM
Visual Feature Extraction
Context-Seed Training Database
Segmentation
Audio Feature Extraction
HMM
Classification with KNN
Training samples of audio sound
Trained audio feature Set
Audio feature sets DB
Speech feature sets DB
Visual feature sets DB
Trained visual feature set
AVCR Unit
Audio Feature Extraction
Audio,Speech Recognition Unit
Video Repository
Context Generator
Phonemes Feature Extraction
Initial Audio-Visual Context
Visual Feature Extraction
Context Matcher
Video Matcher
Query Video clip
Retrieved Videos
Contextual Query Perfection Unit
Browser Display ranked video results
relevant / irrelevant samples
User
CSRF Unit
Spoken Emotional word capture unit
Phonemes feature Extractor
Emotional Word Recognition Unit
Facial Expression capture unit
Facial Expression Recognition Unit
EW
EW
FE
FE
Emotion Recognition Unit
User's Affective Profile Database

Fig. 1. Proposed architecture of AICSRF framework

descriptor and would be used in retrieving the multimedia information from videos. Audio-Visual Context Recognition Unit has four components. First is Audio and Speech Recognizer comprises of Audio Pre-processing Unit, Audio Feature Set Extractor and Phonemes Feature Set Extractor. Second is Visual Feature Set Extractor. Third is Context-seed Training Database. Forth is Context Generator and Matcher. They are discussed in following sections in details.

3.1.1 Audio and Speech Recognizer Unit

Assume that speech signal is realization of some message composed of basic sub-word lexical units, e.g., phonemes or syllables which can be considered as a sequence of symbols. Whole word template are difficult to use for vocabulary recognition, because of whole vocabulary word must be spoken. So, it is preferable to use several smaller descriptive units, such as phonemes, syllables, dyads, etc. in automatic speech recognition system [25]. In this propose paper, phonemes are chosen as sub-word unit. Phonemes list and some spoken words containing these phonemes are given in table 2. The recognition system is trained to recognize the words by breaking them in to sequence of phonemes. Audio and speech recognizer unit is comprises of Audio Processing Unit, Audio Feature Set Extractor, Phoneme Feature Set Extractor . The details are discussed in following sub sections.

Table 2: Phonemes lists

Phonemes : /i:/,/iy:/,/y:/,/uw:/,/uw:/,/ih:/,/uh:/,/ey:/,/er:/,/ow:/,/eh:/,/ah:/,/ao:/,/aa:/,/t:/,/d:/,/n:/,/c:/,/jh:/,/dh:/,/zh:/,/ch:/,/w:/,/a:/,/ai:/,/ae:/,/ei:/,/j:/,/ie:/,/o:/,/oi:/,/th:/,/sh:/,/tz:/,/dz:/,/eu:/,/au:/,/p:/,/m:/,/b:/
Some spoken word containing phonemes: Buy,guy,ajure,die,fie,gin,chin,hang,high,kite,lie,tie,vie,thy,thigh,shy,pie,my,you,zoo,city,marry,mary,merry,new,sang,sane,seen,sin,sky,spy,sty,two,above,bad,bayed,bed,bid,bird,boy,bud,good

3.1.1.1 Audio Feature Set Extractor

Pre-classify into speech and non-speech segment by a KNN classifier and Linear Spectral Pairs-Vector Quantization (LSP-VQ) analysis [26] based on HZCRR, LSTER, SF. From non-speech silence is detected based on STE, ZCR in 1-S window. Calculate BP, NFR, SF and apply rule based music environment and silence discriminating analysis [27] to discriminate music from environment sound.

3.1.1.2 Phoneme Feature Set Extractor

To detect speech boundaries from speech stream, silence and unvoiced frames are removed. STE and ZCR Features are extracted from the frames and used to discriminate silence frames and unvoiced frames. We use Rabiner et. al [28] algorithm based on STE and ZCR for detecting boundaries of speech utterance. The speech signal is now analyzed on a frame by frame basis and the spectral and temporal characteristics of speech are obtained by feature extraction. A frame length of 32 ms with 40 % overlap with previous frame is taken and hamming window is applied to each frame. The first 12 MFCC (static parameter) and 12 MFCC (dynamic parameter) are extracted from each frame to produce 12-D feature vector. We adapt the phoneme-based vector quantization technique given by Yaxin Zhang, et.al. [29] to generate the cluster (codeword). A Gassian Mixture Model (GMM) is estimated for each phoneme by EM algorithm. The EM algorithm [30,31] perform vector quantization over the speech feature vector and generate a codebook in which each codeword is a Gaussian Model having mean vectors, co-variance matrix and a mixture weight. Each cluster in GMM forms a quantization code entry for a phoneme in Phoneme Feature Sets. AdaBoost-HMM classifier [32] comprised of multiple N-state –M symbols HMM adapted as the sub classifier one for discriminating Phonemes Feature Sets and used for word recognition system.

3.1.2 Visual Feature Set Extractor

Background and foreground are separated by applying algorithm given in [33]. Stationary and moving objects are separated frame wise from background and foreground. Total 70-dimensional feature vectors comprising of texture, motion, edge, color, location, background-foreground features are extracted. Six texture features and six motion features are extracted by applying spatiotemporal Gabor filter given in Eq. (1), motion energy given in Eq.(3) and method discussed in section 3.3.3. Nineteen edge features are extracted by applying edge filter and the Water-filling algorithm [34]. Eight shape features are extracted by applying algorithm discussed in [35]. Twenty One color features of image are extracted by applying color histogram. Eight location features and two features for background and foreground are also stored with each feature vector. Thus, sequential combinations of these seven features sets make the 70-dimensional low level feature vectors. Statistical clustering method like k-means or hierarchical clustering [36] is used to classify the visual feature vectors in to number of clusters by mapping Euclidean / Gaussian visual similarity [37].

3.1.3 Context-seed Training Database

In this proposed method multimedia information from videos is retrieved by identifying semantics which depend on context. The context is comprised of audio, speech and visual features present in query clip. To classify the context, good training datasets of audio, speech and visual features which can represent the whole context knowledge are required. Increasing the numbers of good training samples increases the discriminating power of classifiers. It is tedious, cumbersome yet impossible to collect sufficient and good training samples of audios, speech and visual objects and train the classifier to discriminate whole context accurately. However, we have developed initial Context-seed Training Database having trained Audio Feature sets DB, trained Speech Feature Sets DB and Visual Feature Sets DB. For this purpose, we have recorded sound of 4000 spoken words uttered by 50 subjects (male and female) of various personal, familial or cultural traits iterated ten times for speech training samples sets. We have downloaded sound files of 500 music clips from internet for audio training samples sets and 50 images (of different size, shape, orientation and morphed) corresponding to each spoken words and audios for visual training samples sets. Audio Speech Recognizer method applied in section 3.1 is applied to extract the phoneme feature sets from the speech training samples set. HMM [38] is applied to train and classify the spoken word feature sets. This trained spoken word feature sets are stored in the Trained Speech Feature Sets DB. Similarly audio feature sets are extracted from the audio training samples set, HMM is applied to train and classify the audio feature sets and the trained audio feature sets are stored in the Trained Audio Feature Sets DB. Visual features (shape, edge, texture, color) are extracted from training images samples set. Image is segmented classified by clustering or using KNN classifier [39] and resultant trained visual feature sets are stored in Trained Visual Feature Sets DB. This initial Context-seed Training Database having Trained Audio Feature Sets DB, Trained Speech Feature Sets DB and Trained Visual Feature Sets DB would be used with Context Matcher in AVCRU to match audio-visual context.

3.1.4 Context Generator and Context Matcher

Audio and speech recognized by Audio Speech Recognizer Unit provides the primary context and a visual objects recognized by Visual Object Recognizer provides the secondary context. The composite audio, speech and visual is the context in which user wants to retrieve the information. The audio and speech extracted from the query clip is matched with the Audio Feature sets DB and Speech Feature Sets DB in Context-seed Training Database by Cosine similarity [40]. Any match founds, becomes the audio context. After finding audio context, the visuals present in query clip is matched with Trained Visual Feature Sets DB in Context-seed Training Database by Cosine similarity. If match found, the matched visual becomes the visual context. Combining these matched audio contexts and visual contexts multiple Audio-Visual contexts are generated. Four possible combinations of contexts are generated. These are (i) only audio context present (ii) only visual context present (iii) both audio and visual context present (iv) both audio and visual context absent. These initial contexts will be used to retrieve the videos from repository by matching audio context first then visual context. We will study the retrieval performance in all the possible combinations of Audio-Visual contexts.

3.2 Contextual Semantic Relevance Feedback

RF is a powerful tool traditionally used in text based information retrieval system [41]. The RF in CBIR is used to bring the user in relevance loop to fill the semantic gap between low-level features and what a user perceives in his mind that is high level concepts. The steps for RF in CBIR proceed as in first step, the system provides the initial retrieved results through query by examples, sketch, etc. In second step, a user judges the above results as relevant and irrelevant to the query and in third step, machine learning algorithm is applied to learn the user's feedback then repeat second step and third step until user satisfied with the results [42]. The typical approaches in third step are 're-weighting' dynamically update the weights embedded in query [43,44], query-point-movement QPM [44,45,46] and SVM is often used to capture a query concepts by separating the relevant images from irrelevant images using hyper-plane in projected space [44,47,48,49]. In most of these RF systems, keeping the similarity measurements fixed, the importance or weight of each descriptor is estimated through the RF from user. They uses only the low level features to estimate the ideal query parameters and don not address the 'semantic' of images. To circumvent the problem of learning from small training sets, user's patience in supporting multi-round feedback, semantic concept variability of user's perception in various contexts under which user wants to retrieve the information, the proposed relevance feedback system categorize the relevant and irrelevant contents by considering the contextual semantics. The proposed Contextual Semantic Relevance Feedback proceeds under the following steps:

Step 1: The Audio-Visual Context Generator Recognition Unit recognizes the Contexts. The initial retrieved results are displayed in Browser by matching Audio-Visual features present in a query clip under a particular context to the Audio -Visual features present in the videos of the video repository under the same context.
Step 2: To categorize the above results as whether and to what degree, they are relevant and irrelevant videos under context, ERU provides affective feedback by capturing emotions when a user opens videos to see the desired contents.
Step 3: Contextual Query Perfection Scheme discussed in section 3.4, learns the contextual semantics and refines the context by taking user's emotion as feedback, then go back to step 2. Iterate loop of Step 2 and Step 3 until user satisfied with the results.

3.3 Emotion Recognition Unit (ERU)

One of our goals is to model user's affective behavior and to identify a rich set of affective features that will be used in determining relevant and irrelevant information as per user's contextual perceptions and need. According to emotion theorists, six or more basic emotions (sadness, happiness, anger, fear, surprise, disgust) exists [50]. In [51], authors indicate that emotions are communicated through facial expressions. Facial expressions as conversational signals, which show not only the focus of attention [52], but also provide useful cues (smiles, chuckles, smirks, frowns, etc.) that are considered an essential part of our social interaction [20].

As emotion is subjective, some user may use emotional words along with facial expressions during the information retrieval process while others may not. Our assumption is that, if a user, shows emotions by combination of spoken emotional words and facial expressions during the information retrieval process, this can be the effective evidence to feedback cycle to predict the reasonable accuracy of relevant and irrelevant information. Our model incorporates both spoken emotional words and facial expressions as affective features.

3.3.1 Affective profile database of user

Several attempts have been introduced in the past to build user's profile and learning user interest from implicit feedback [53,7]. Constructed user profile is based on user's search history, web pages visited, documents created and viewed, etc, and lead effectively to more accurate relevance prediction [54]. In our model, for contextual semantic relevance, we incorporate user's affective profile database which store the spoken emotional words and real emotion captured during the information retrieval process. When recording facial expressions several critical issues arise [55]. Emotional expressions are highly idiosyncratic in nature and may vary significantly from one individual to another (depending on personal, familial or cultural traits). In our approach we employed a facial expression recognition system of reasonably robust performance and accuracy across all individuals by keeping real affective profile database, capturing real facial expressions and spoken emotional words, thus increasing the chance of observing real spontaneous behaviour during retrieval process. So we propose a method in our paper to obtain good real training dataset of emotions shown by facial expressions and emotional spoken word. Initially, store audio spoken with corresponding facial expression in user profile. During the retrieval process, actual expressions are captured and previously stored facial expressions are refined in user profile. The new refined facial expression database with spoken emotional words becomes the new user emotional profile database and is used as training datasets.

3.3.2 Emotional spoken word recognition System

One objective is to design an emotional spoken word recognition system with small emotional words vocabulary to evaluate the effectiveness of acoustic speech signals in detecting spontaneous changes in emotion during multimedia retrieval process. The small emotional spoken words vocabulary consists of emotional words grouped in categories to express the six basic emotions. Acoustic speech sound (phonetic) may vary significantly from one individual to another (depending on personal, familial or cultural traits and on their vocal articulation) and it is difficult to segment and label transitional part of speech. Considering above factors, we propose an emotional spoken word recognition system of reasonably robust performance and accuracy across all individual speakers by keeping user's real affective profile database and capturing and storing real spontaneous sound of spoken emotional words. Initially utterance of emotional words is classified as a certain phoneme or set of phonemes and is stored in user profile. During the retrieval process, actual sound of emotional word is captured, phonemes feature sets are extracted from the spoken emotional words. These phonemes feature sets are actual representative training feature sets of six basic emotion categories for a particular user and stored in User's Affective Profile Database.

Our proposed Emotional Spoken Word Recognition system uses the phonemes produced during the articulation phase (the period when the sound is produced in utterance) as the sub-word units in recognizing the spoken emotional words. By modeling and recognizing phonemes, the system may be further trained to recognize words by breaking them in to sequence of phonemes. Acoustic phonemes feature sets extracted from audio signal are used for classification of emotional words. The multi-dimensional distributions of phonemes are a Gaussian like hyper ellipsoidal structure [30]. The phonemes are treated as cluster in speech signal space. The method applied in section 3.1.1.2 is used to extract phonemes feature sets. The success of modeling and analysis of temporal processes by HMM (Hidden Morkov Model) is well reported in literature [56, 57]. Most of the techniques take a word as the basic unit for recognition. This work well for small vocabulary but large vocabulary problem can be solved by choosing phonemes as sub-word unit. Recent advance techniques have been using couple HMM, product HMM and factorial HMM for audio visual in training and recognition of emotion [58]. In this proposed paper phonemes are chosen as sub-word units. The recognition system is trained to recognize words by breaking them in to sequence of phonemes. In this proposed recognition system, AdaBoost-HMM classifier comprised of multiple N-state –M symbols HMM adapted as the classifier for discriminating Phonemes Feature Sets and is used in training and recognizing the spoken emotional word.

3.3.3 Facial features extraction and recognition system

Facial expression is recognized by capturing spontaneous changes in facial evidences such as eye region, the gap between eyebrows, forehead, region around nostrils, the corners of mouth, chick, etc. Facial expression recognition system, to be used in multimedia retrieval process, requires robust and fast detection of real spontaneous changes in facial feature. Our plan is to use texture features and motion for tracking and measuring any spontaneous facial deformation. The regions of interest in human faces are comprise of concave (eye, the gap between lips and chin, area around nostrils and nose, etc.) and convex (forehead, chick, eyebrows region, lips, etc.) having minimum intensity and maximum intensity respectively. Any spontaneous deformation in these concave (local minima) and convex (local maxima) regions would be tracked and measured by applying motion energy given by Eq.(3). These motion energy units obtained are matched with the motion energy units of real emotional training datasets of users in their affective profile database, finally classified into basic emotion categories. Computational model for feature extraction and motion analysis is as follows:

In seminal work, Adelson and Bergen [59] suggested that a two-dimensional spatial pattern moving at a given velocity corresponds to a three-dimensional spatiotemporal pattern of a given orientation which can be detected with an appropriately oriented 3D spatiotemporal filter, such as a 3D Gabor filter. In [60] Nikolai Petkov , et al., model the spatio temporal receptive field profile of simple cells as a family of Gabor filter function denoted by $g_{v,\theta,\varphi}(x, y, t)$ where $(x,y,t) \in \Omega \subset R^3$ which is centered in the origin (0, 0, 0) as given in Eq.(1).

$$g_{v,\theta,\varphi}(x,y,t) = \frac{\Upsilon}{2\pi\sigma^2}\exp\left(\frac{-((\bar{x}+v_c t)^2+\Upsilon^2\bar{y}^2)}{2\sigma^2}\right)$$
$$.\cos\left(\frac{2\pi}{\lambda}(\bar{x}+vt)+\varphi\right)$$
$$.\frac{1}{\sqrt{2\pi}\tau}\exp\left(-\frac{(t-\mu_t)^2}{2\tau^2}\right).U(t) \quad (1)$$

Where

$$\bar{x} = x\cos(\theta)+y\sin(\theta)$$
$$\bar{y} = -x\sin(\theta)+y\cos(\theta)$$
$$U(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{if } t \leq 0 \end{cases}$$

Here, the parameter v is the preferred speed, the angle parameter θ determines the preferred direction of motion and the preferred spatial orientation of the filter, and φ is a parameter that determines the spatial symmetry of the function. The phase offset φ ∈ (−π, π) determines the symmetry of $g_{v,\theta,\varphi}(x, y, t)$ in the spatial domain. The standard deviation σ of this Gaussian factor determines the size of the receptive field. Other parameters detail may be found in Nikolai Petkov, et al. [60]. $g_{v,\theta,\varphi}(x, y, t)$ is a product of a Gaussian envelope function that restricts $g_{v,\theta,\varphi}(x, y, t)$ in the spatial domain, a cosine wave traveling with a phase speed v in direction θ , another Gaussian function, with a mean μ_t and standard deviation τ. Authors further investigated the direction and speed tuning properties of the motion energy filter. The maximum response of filter is obtained when the preferred direction of the filter (θ) matches the direction of movement of stimuli. Filter that prefers higher speed of travelling cosine wave have bigger receptive field. The response $r_{v,\theta,\varphi}(x, y, t)$ of a linear filter with a RF function $g_{v,\theta,\varphi}(x, y, t)$ to a luminance distribution l(x, y, t) is computed by convolution as follows:

$$r_{v,\theta,\varphi}(x,y,t) = I(x,y,t) * g_{v,\theta,\varphi}(x,y,t) \quad (2)$$

3.3.3.1 Motion detection

In case of spatiotemporal filter, input from the present and the output from the past is used to compute the response at the current moment and can be used for analyzing and detecting motion. Motion at a given spatial position can be detected in straightforward way from the motion energy equation. Motion energy equation is the phase sensitive response obtained by quadrature pair summation of the responses of two filters with a phase difference of $\pi/2$ given below:

$$M_{v,\theta,\varphi}(x,y,t)= \sqrt{(r^2_{v,\theta,\varphi}}(x,y,t) + r^2_{v,\theta,\pi/2}(x,y,t)) \quad (3)$$

3.4 Contextual Query Perfection Scheme

Semantic meaning often changes with context. Single context or multiple contexts may be present in videos. The AVCR unit can provide single distinct context and multiple contexts either distinct or co-occurring simultaneously. In this proposed work, we are confined only presence of single distinct contexts in a query. The issues and challenges arise due to multiple contexts co-occurring simultaneously in a user's query will be discussed and tackled in my future work. Contextual semantic knowledge represents the relationship between visual concepts under certain context. As we have discussed the audio and speech are the primary context descriptor, visual is the secondary context descriptor and combination of these two context descriptor is the audio-visual context descriptor. The Contextual Query Perfection Scheme will evaluate the contextual fitness of semantic concept and refine audio-visual context on the basis of relevance judgment provides by ERU via relevance feedback. One of the most important and crucial step in CSRF is the third step that is machine learning algorithm to learn the contextual semantic and can refine the context by taking user emotions as feedback. The system dynamically learns the user's intention, and gradually present better search results. Query point movement and dynamically re-weighting the feature weight embedded in query are two widely used approaches. To evaluate the proposed framework we are using two state-of-the-arts algorithms for baseline algorithm. One of them is Rocchio's Algorithm [61] and second is scheme discussed in MindReader System [62]. Contextual Query Perfection Scheme uses these algorithms to improve the estimation of query point by moving it towards positive samples and away from the negative samples.

Let $PV^+ = \{V_i \mid i = 1,2,3,\ldots\alpha\}$,represent the set of +ve videos samples namely Positive Video Bag and $NU^- = \{U_j \mid j = 1,2,3,\ldots,\beta\}$, represent the set of -ve videos samples namely Negative Video Bag categorized by the system based on user's positive and negative emotion captured by ERU respectively. The system categorizes the Sub-Bag of +ve videos from the Positive Video Bag and Sub-Bag of –ve videos from the Negative Video Bag according to contexts present. The system identifies the videos V_ℓ (for some ℓ, $1 \le \ell \le \alpha$) in PV^+ having context c_i, and the set $A_{c_i V_\ell}$ = $\{ V_\ell^{c_i} = V_\ell$, for some ℓ, $1 \le \ell \le \alpha \}$ with $\left|A_{c_i V_\ell}\right|$ = n (say) $\le \alpha$ that represents the set of +ve videos samples under context c_i. The | . | denotes the cardinality of set i.e. the number of elements in set. The system also identifies the videos Ut (for some t, $1 \le t \le \beta$) in NU^- having context c_i, and the set $B_{c_i U_t}$ = $\{ U_t^{c_i} = U_t$, for some t, $1 \le t \le \beta \}$ with $\left|B_{c_i U_t}\right|$ = m (say) $\le \beta$ represents the set of -ve videos samples under context c_i. The context c_i may be primary context denoted by c_i^p or may be secondary context denoted by c_i^s. In case of single distinct context identified, all the +ve videos in the set PV^+ and all the –ve videos in the set NU^- would be used in finding similarity measure based on either shortest Euclidean distance or highest Gaussian probability measure. The context refinement target would be to maximize the visual feature sets similarity for the relevant samples, i.e. +ve video samples, and the same time to minimize the visual feature sets similarity for irrelevant, i.e. –ve video samples. We modify both algorithms as follows:

3.4.1 Contextual Query Perfection Scheme A

This scheme is based on method discussed in "MindReader System" [62] for relevance feedback. It improves the estimation of the query point by moving it towards the positive samples from the new query. This is the integration of two approaches (Query Point Movement and feature re-weighting) into a unified framework, based on the minimization of total distances of positive samples from the new query. We have modified and applied it to work under context and the optimal query vector can be adaptively approached as follows:

Let α be the number of positive videos and let

$V_{\ell}^{c_i} = [v_{\ell 1}^{c_i}, \ldots, v_{\ell d}^{c_i}]^T$ be a d-dimensional vector that represents ith –video $(\ell = 1, \ldots, \alpha)$ under context c_i.

Let X^{c_i} denotes an $\alpha \times d$ matrix $X^{c_i} = [V_{\ell}^{c_i}, \ldots, V_{\ell\alpha}^{c_i}]^T$. Vector $V^{c_i} = [v_1^{c_i}, \ldots, v_{\ell}^{c_i}]^T$ represents the degree of relevance for the ℓ relevant images given by the user under some context c_i

The distance function is given by

$$D(V_{\ell}^{c_i}, q) = (V_{\ell}^{c_i} - q)^T M (V_{\ell}^{c_i} - q) \qquad (4)$$

The optimal solution to **q** and **M** are as follows $q = \frac{X^{c_i} V^{c_i}}{\sum_{i=1}^{\alpha} V_{\ell}}$

$$, \ M = (\det(C))^{\frac{1}{d}} C^{-1} \qquad (5)$$

Where C is the weighted covariance matrix, whose elements are computes in the following way:

$$C_{mn} = \sum_{\ell=1}^{\alpha} V_{\ell}^{c_i} (V_{\ell n} - q_n)(V_{\ell m} - q_m) \qquad (6)$$

The optimal query q in (5) turns out to be the weighted average of the relevant images under context. The symmetric full matrix M enables the system to estimate diagonal queries under certain context.

3.4.2 Contextual Query Perfection Scheme B

This scheme is based on Rocchio's formula[61] for relevance feedback. It improves the estimation of the query point by moving it towards the positive samples away from negative samples. The optimal query vector can be adaptively approached by Rocchio's formula

$$Q_{new} = \alpha Q_{old} + \beta\left(\frac{1}{N_{R'}} \sum_{i = D'_R} D_i\right) - \gamma\left(\frac{1}{N_{N'}} \sum_{i = D'_N} D_i\right) \qquad (7)$$

Where Q_{old} and Q_{new} are the original and updated query, respectively, D'_R and D'_N are the positive and negative samples returned by the user, $N_{R'}$ and $N_{N'}$ are the number of samples in D'_R and D'_N, respectively, and α, β, γ are selected constants. Our proposed Contextual Query Perfection Scheme A is modified version of rocchio's formula to estimate the query point by moving it towards the positive samples away from negative samples under a particular context hence new formula for optimal query vector under context is now given by

$$Q_{new}^{c_i} = \alpha Q_{old}^{c_i} + \beta\left(\frac{1}{N_{R'}^{c_i}} \sum_{i = D'^{ci}_R} D_i^{c_i}\right) - \gamma\left(\frac{1}{N_{N'}^{c_i}} \sum_{i = D'^{ci}_N} D_i^{c_i}\right) \qquad (8)$$

Where $Q_{old}^{c_i}$ and $Q_{new}^{c_i}$ are the original and updated query, respectively, $D'^{c_i}_R$ and $D'^{c_i}_N$ are the positive and negative samples returned by the user under context c_i, $N_{R'}^{c_i}$ and $N_{N'}^{c_i}$ are the number of samples in $D'^{c_i}_R$ and $D'^{c_i}_N$, respectively, and α, β, γ are selected constants. In our proposed work α, β, γ all are having value 1. The values of $D'^{c_i}_R$ and $D'^{c_i}_N$ are as $D'^{c_i}_R = |A_{c_i} V_{\ell}|$ = n (say) that represents the set of +ve videos samples under context c_i that is $V_{\ell}^{c_i}$ and $D'^{c_i}_N = |B_{c_i} U_t|$ = m (say) represent the set of -ve videos samples under context c_i that is $U_t^{c_i}$. The contexts from the Contextual Query Perfection scheme are again used to retrieve the videos. The process is repeated until user satisfied with the result sets. The visual content associated with set of contexts to which they are related is stored in Contextual Semantic Database.

4. Experimental setup, results and discussion

We started from collection of the development dataset for video repository. The video repository contains a large number of videos provided by open-video project [63]. The video collection includes TRECVID 2001 and TRECVID 2002 videos datasets containing NASA Educational, Historical, Ephemeral, Miscellaneous, Internet Moving Archive, National Archive, Stillmans Fire Collections, Informedia project at Carnegie Mellon University categories downloaded manually. We have also obtained video database CC_WEB_VIDEO comprise of duplicate and near duplicate videos provided by Video Retrieval Group (VIREO) city of Hong Kong and Informedia Group from Carnegie Mellon University [64]. Silent category videos was also downloaded from open-video project website to study the effect of occurrence of Secondary context in contextual semantic retrieval. Video repository also includes lecture videos downloaded from Internet, videos from Discovery channel, Planet Animal channel, news videos captured from CNN, MSNBC. The development dataset of videos in video repository used in this experiment contain 4874 videos. Video clips dataset was also prepared manually by separating video clips of 3s, 5s duration from the videos of repository and nearly 500 hundred music video clips were also collected from Internet. Affective Profile training database was prepared during enrollment of users. User utters a list of emotional words. During this utterance of emotional words facial expression is also captured by webcam with built in microphone. Each emotional word is uttered 20 times. Audio signal was separated from video of face and stored in .wav sound file. Facial expression video was stored in .avi format. Thus 20 audio and 20 video files are stored for each user in their affective profile training dataset.

After the collection of database a complete experimental setup and algorithms was implemented on intranet and client server architecture was used for video retrieval. Video repository, user affective profile database and contextual seed training database was stored in server running MATLAB and MSP (MATLAB Server Page). All client computers having 2.53 GHz processor, 2 GB RAM equipped with web-cam and built in microphone to capture emotional words and facial expression during video retrieval process. We evaluated our AICSRF framework introduced in section 3. ERU unit captures emotions shown on contents, and discriminates a user's emotion into positive and negative emotion. The system categorizes the retrieved video result sets into positive and negative sample sets based on positive and negative emotion recognized by ERU. The system again categorizes the positive and negative sample sets into Sub-Bag of positive videos and Sub-Bag of negative videos according to the contexts present in the query clip. At every feedback round the system's Contextual Query Perfection Scheme estimate the new query point and retrieve the videos. We followed each relevance feedback session for 5 iterations and measured the precision, recall and average retrieval time.

4.1 Study of the effect of probability of occurrence of contexts in contextual semantics relevance feedback

We studied the effect of context in video information retrieval and contextual semantic understanding of system by exhaustive retrieval testing. The proposed AICSRF framework was tested first without considering context, secondly considering primary context (audio) only, thirdly considering secondary context (visual) only and finally considering joint context (audio + visual). We built precision vs. iteration diagram. The precision of our proposed scheme, which combines audio, speech and visual as joint context, is shown in Fig. 2. These observations can be easily found. The combination of the audio, speech with visual as joint context consistently improves the understanding of contextual semantics. In specific, it improves the precision from 58% to 82% for our proposed scheme that uses joint context in comparison to secondary context (visual) only. The addition of the audio, speech with visual stabilizes the semantic fluctuation associated with visual context based retrieval, which are mainly caused by the inaccurate classification and ambiguity in semantics under a context, that is, the difference between the highest and lowest precision is 24% and 32% for our proposed scheme (joint context based) and secondary context (visual) based scheme respectively. We see that joint use of audio and visual context signature provide significant improvement of contextual semantic concepts as compared to the use of visual context signature based retrieval. We also studied the effect of context on retrieval time and search space. We measured the retrieval time vs. iteration and built retrieval time vs. iteration diagram that is shown in Fig. 3. In specific, the shortest vs. longest retrieval time is 0.52 minutes against 1.68 minutes. However, the corresponding precision is 82% and 65%. Our proposed scheme takes about 6.45 minutes retrieval time in five iterations and yields 82% precision in contextual semantic understanding. We found that joint use of audio and visual as context signature had taken less retrieval time in giving the video result sets according to user's interest. This validates our proposed scheme and best compromises in terms of precision and efficiency for retrieval of video according to a user interest from large video repository.

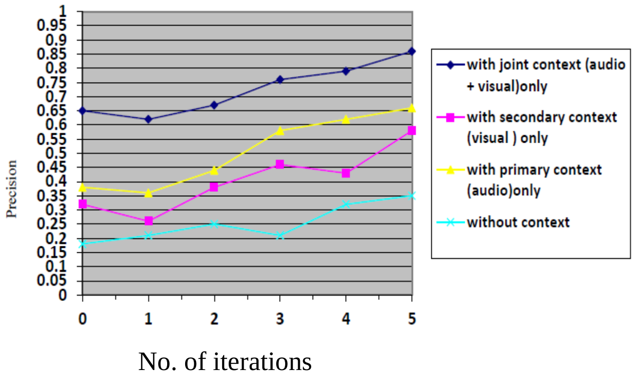

Fig. 2. Precision comparison on considering occurrence of primary context only, secondary context only, joint context only and without context.

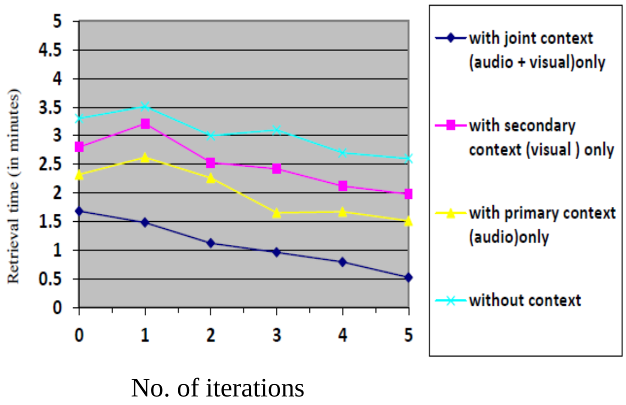

Fig. 3. Retrieval time considering occurrence of primary context only, secondary context only, joint context only and without context.

4.2 Study of the effect of affective feedback as source of evidences in contextual semantic relevance feedback

We evaluated the AICSRF framework with applying emotional recognition techniques proposed in ERU unit and studied the effect of occurrence of these emotions in contextual semantic relevance feedback. We followed up the iterations by applying first emotional words (spoken emotional word) only as source of evidences in RF, second by applying facial expression only as source of evidences in RF and in last by applying joint (emotional spoken words + facial expression) as source of evidences in RF with considering primary context, secondary context, joint context and without context. We measured the accuracy, precision and recall of video retrieval for each case. The results are shown in Table 3., Table 4. And Table 5., respectively. The model that trained on joint (spoken emotional word + facial expression) as source of evidences in RF held the best accuracy and out performed in both precision (82.2%) as well as in retrieval time (6.45 in five iterations) than other emotional recognition techniques discussed in ERU unit. The precision vs. number of iterations and retrieval time vs. number of iterations diagrams considering proposed emotional techniques are shown in Fig. 4 and Fig. 5 respectively. The result supports our hypotheses that combination of spoken emotional words and facial expressions during the information retrieval process would be the effective evidence to feedback cycle to predict the reasonable accuracy of relevant and irrelevant information.

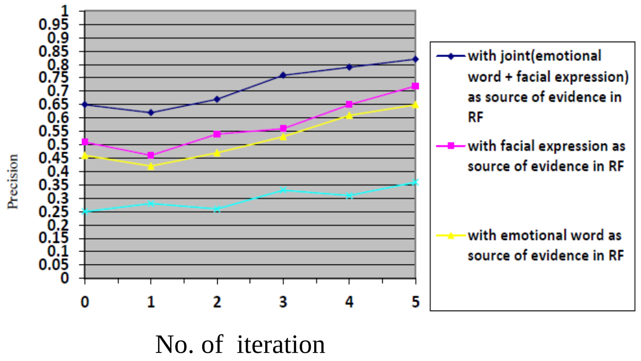

Fig. 4. The effect of joint (spoken emotional word + facial expression) as source of evidence in RF, spoken emotional word as source of evidence, facial expression as source of evidence and without affective feedback in RF on Precision.

Table 3: considering emotional words (spoken emotional word) as source of evidences in RF

	Accuracy (%)	Precision (%)	Recall (%)
With primary context only	50.8	50.4	66.7
With secondary context only	56.4	55.8	59.2
With joint context only	65.6	65.2	66.8
Without context	48.0	46.6	47.8

Table 4: considering facial expression as source of evidences in RF

	Accuracy (%)	Precision (%)	Recall (%)
With primary context only	54.6	54.2	62.8
With secondary context only	58.8	58.6	59.2
With joint context only	72.4	72.2	70.2
Without context	54.0	53.8	57.6

Table 5: considering joint (spoken emotional word + facial expression) as source of evidences in RF

	Accuracy (%)	Precision (%)	Recall (%)
With primary context only	60.8	60.1	58.6
With secondary context only	66.2	65.8	65.4
With joint context only	82.6	82.2	80.2
Without context	56.4	56.0	55.2

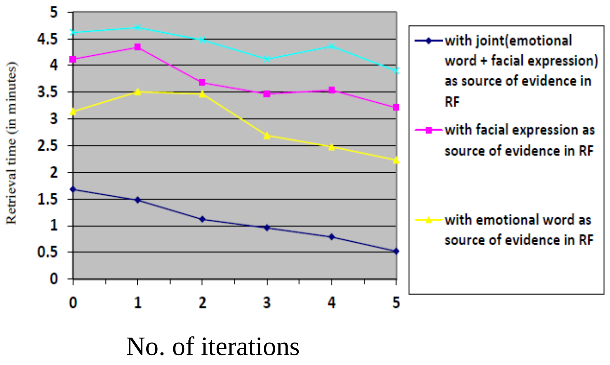

Fig. 5. The effect of joint (spoken emotional word + facial expression) as source of evidence in RF, spoken emotional word as source of evidence, facial expression as source of evidence and without affective feedback in RF on retrieval time.

4.3 Comparative Study of the Contextual Query Perfection schemes in contextual semantic relevance feedback

To show the effectiveness of the Contextual Query Perfection Scheme, we have simulated user's emotional feedback and evaluated our AICSRF as follows. For a query video clip, in each iteration, the system categorizes the emotions of a user by ERU and examines top 40 videos that are more similar to the optimal query. The Contextual Query Perfection Scheme A uses only positive video samples for the estimation of query point in next iteration while Contextual Query Perfection Scheme B uses both positive videos and negative videos samples for the purpose. The Precision and Retrieval time of both schemes are compared and shown in Fig. 6 and Fig. 7, respectively. The results of the two query sets show that, after five iterations of feedback, the Precision of Contextual Query Perfection Scheme B is higher than that of Contextual Query Perfection Scheme A, i.e., 82% and 64%, respectively. The superiority of Contextual Query Perfection Scheme B over Contextual Query Perfection A is probably due to more discriminating power of Contextual Query Perfection Scheme B to identify the context of a user query that reduce the search space. Due to reduced search space in next iteration Contextual Query Perfection Scheme B takes less retrieval time in compare to Contextual Query Perfection Scheme A that is 0.51 minutes and 1.51 minutes, respectively.

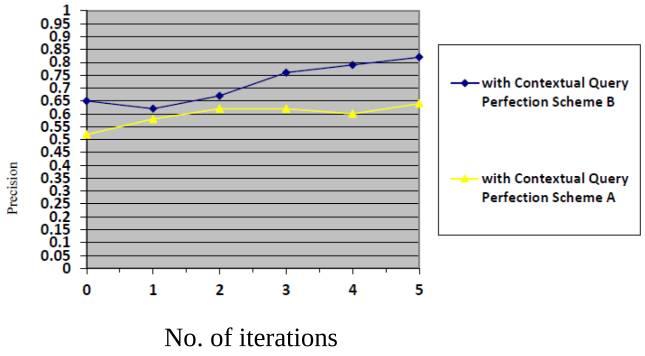

Fig 6: Precision results of Contextual Query Perfection Scheme A and Contextual Query Perfection Scheme B with joint (spoken emotional word + facial expression) as source of evidence in RF

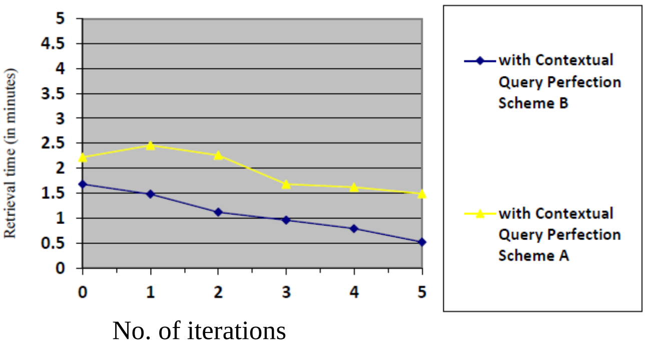

Fig 7: Retrieval time comparison of Contextual Query Perfection Scheme A and Contextual Query Perfection Scheme B with joint (spoken emotional word + facial expression) as source of evidence in RF

5. Concluding remarks and future directions

In this paper, we have proposed an AICSRF framework to investigate whether audio and speech could determine the context under which a user wants to retrieve the information and to further investigate whether we could employ affective feedback as implicit source of evidence in CSRF cycle to increase the system's contextual semantic understanding. We have introduced *Contextual Query Perfection Scheme* to learn, refine the current context that could be used in query perfection in RF cycle to understand the semantic of query in following iteration. We have also put forward an improved RF method employing a new selection criteria based on emotional words along with facial expression as source of evidence in RF to reduce the ambiguities in semantics under multiple contexts. We have introduced an ERU unit in framework that comprises of a customized 3D spatiotemporal Gabor filter to capture spontaneous facial expression by detecting motion in 3-dimention and emotional word recognition system that uses phonemes to recognize the spoken emotional words. The result shows that this emotion recognition technique has more discriminating characteristics in selecting relevant and irrelevant information. The feasibility of AICSRF framework was demonstrated for contextual semantic understanding. The framework was validated by testing first without considering context, secondly considering primary context (audio and speech) only, thirdly considering secondary context (visual) only, and finally considering joint context (audio, speech and visual). As shown by experimental evaluation performed on a large video repository, the use of audio and speech with visual feature sets as joint context along with emotional words and facial expression as source of evidence in our proposed AICSRF framework contributed a significant improvement in contextual semantic understanding of the system. The test results support our both hypotheses that audio and speech in determining the context under which a user wants to retrieve information would increase the contextual semantic understanding and combination of spoken emotional words with facial expression would be effective evidence to feedback cycle in selecting relevant and irrelevant information. The system is working fine but still requires a technique that can discriminate emotions more precisely so that framework can learn, refine, discriminate the current context in the user query hence can reduce the search space and retrieval time with high precision. Another companion work of ours' will demonstrate the study of the effect of addition of discriminating power of classifier algorithm along with AICSRF on the retrieval performance and further study the effect of increasing the Affective Features in identifying, learning, refining and discriminating emotions more precisely on overall retrieval performance. We will also study the effect of increasing the size of good training samples of initial contexts and whereby we will try to find the answer of the interesting question "How many training sample sets of audio, speech and visual features would be sufficient to recognize the whole context and fill the semantic gap?".

References

[1] M. S. Kankanhalli and Y. Rui, "Application Potential of Multimedia Information Retrieval", Proceedings of the IEEE, 96 (4),2008.

[2] E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior information", In SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2006, pp. pages 19-26.

[3] R. Badi, S. Bae, J. M. Moore, K. Meintanis, A. Zacchi, H. Hsieh, F. Shipman and C. C. Marshall, 'Recognizing user interest and document value from reading and organizing activities in document triage", In Proceedings of the 11th international conference on Intelligent User Interfaces, ACM, 2006, pp. 218-225.

[4] D. R. Hardoon, J. S.Taylor, A. Ajanki, K. P. aki and S. Kaski, "Information retrieval by inferring implicit queries from eye movements", In Eleventh International Conference on Artificial Intelligence and Statistics, 2007.

[5] J. Koenemann and N. J. Belkin, "A case for interaction: a study of interactive information retrieval behavior and effectiveness", In CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM,1996, pp. 205-212.

[6] Y. Rui and S. Huang, "Optimizing learning in image retrieval", In IEEE Proceedings of Conference on Computer Vision, 2000, pp. 236-243.

[7] K. Puolamaki, J. Salojarvi, E. Savia, J. Simola and S. Kaski, "Combining eye movements and collaborative filtering for proactive information retrieval", In SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM ,2005,pp. 146-153.

[8] Y. Aytar,O. B. Orhan and M. Shah, " Improving semantic concept detection and retrieval using contextual estimates", ICME, 2007.

[9] T.Yoshizawa and H. Schweitzer, "Long-term learning of semantic grouping from relevance-feedback", In ACM SIGMM International Workshop on Multimedia information retrieval, ACM,2004, pp. 165–172.

[10] F. Hopfgartner, "A news video retrieval framework for the study of implicit relevance feedback", In Proceedings of the Second International Workshop on Semantic Media Adaptation and Personalization, IEEE Computer Society,2007, pp. 233-236.

[11] C. A. Chin, A. Barreto, J.G. Cremades and M. Adjouadi, "Performance analysis of an integrated eye gaze tracking / electromyogram cursor control system", In Proc. 9th Int. ACM SIFACCESS Conf. on Computers and accessibility,2007, pp. 233–234.

[12] D. K. Limbu, A. Connor, R. Pears and S. MacDonell, **"**Contextual Relevance Feedback in Web Information Retrieval", Information Interaction in Context, ACM,2006, pp. 138-143.

[13] J.S. Park, K.B. Eum, K.H. Shin and J.W. Lee, "Color Image Retrieval Using Emotional Adjectives", Korea Information Processing Society, B, 10-B (2),2003, pp.179-188.

[14] E.J. Park and J.W. Lee, "Emotion-Based Image Retrieval Using Multiple-Queries and Consistency Feedback", The IEEE National Conference on Industrial informatics, India, 2008.

[15] J. Salojarvi, K. Puolamaki and S. Kaski, "Implicit Relevance Feedback from Eye Movements", Artificial Neural Networks: Biological Inspirations ICANN 2005, Springer, 3696 (2005).

[16] I. Arapakis, I. Konstas and J. M. Jose, " Using Facial Expressions and Peripheral Physiological Signals as Implicit Indicators of Topical Relevance", In SIGIR '09:Proceedings of the 32st annual international conference on Research and development in information retrieval, ACM, 2009.

[17] I. Arapakis, Y. Moshfeghi, H. Joho, R. Ren, D. Hannah and J. M. Jose, " Enriching user proffling with affective features for the improvement of a multimodal recommender system", In Conference on Image and Video Retrieval, 2009.

[18] P. Ekman, "Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life", Times Books, 2003.

[19] M. Pantic and L. Rothkrantz, "Expert system for automatic analysis of facial expression", Image and Vision Computing Journal, 18(11), 2000, pp. 881-905.

[20] A. Russell, J. J. Bachorowski and J. F. Dols, "Facial and vocal expressions of emotion", Annual Review of Psychology, 2003.

[21] A. Salway and M. Graham, "Extracting information about emotions in films", In: Proceedings of ACM Multimedia '03, 2003.

[22] H. Miyamori, S. Nakamura and K. Tanaka, " Generation of views of TV content using TV viewers' perspectives expressed in live chats on the web", In: Proceedings of ACM Multimedia '05, 2005, pp. 853–861.

[23] C.H. Chan and G.J.F. Jones, " Affect-based indexing and retrieval of films", In: Proceedings of ACM Multimedia '05, 2005, pp. 427–430.

[24] Y.J. Chang, C.K. Heish, P.W. Hsu and Y.C. Chen, "Speech-Assisted Facial Expression Analysis and Synthesis for Visual Conferencing System", Proceedings of ICME, 2003, pp. 111 – 529.

[25] S. Hayamizu, K.Tanaka and K. Ohta, "A Large Vocabulary Word Recognition System Using rule based Network Representation of Acoustic Characteristic Variations", IEEE,1988.

[26] J. P. Campbell, "Speaker recognition: A tutorial", Proc. IEEE, 85 (9),1997, pp. 1437–1462.

[27] L. Lu, H.J. Zhang and H. Ziang, "Content Analysis for Audio Classification and Segmentation", IEEE Transactions on Speech and audio Processing, 10 (7) 2002, pp. 505-515.

[28] L. Rabiner, et al., “Fundamentals of Speech Recognition”, Prentice Hall, 1993, pp. 45-50..

[29] Y. Zhang, R. Togneri and M. Alder, “Phoneme-Based Vector Quantization in a Discrete HMM Speech Recognizer”, IEEE Transactions on Speech and Audio Processing, 5 (1), 1997, pp. 26-32.

[30] J. H. Wolf, “Pattern clustering by multivariate mixture analysis”, Multivariate Behav, Res., 5, 1970, pp. 329–350.

[31] P. McKenzie and M. Alder, “ Initializing the EM algorithm for use in Gaussian mixture modeling”, in Proc. Pattern Recognition, 1994.

[32] S. W. Foo, Y. Lian and L. Dong, “Recognition of Visual Speech Elements Using Adaptively Boosted Hidden Markov Models”, IEEE Transactions on Circuits and Systems for Video Technology, 14 (5), 2004, pp. 693-705.

[33] B. GuKnsel, A.M. Ferman and A.M. Tekalp, “Temporal video segmentation using unsupervised clustering and semantic object tracking”, J. Electron. Imag., **7** **(**3),1998, pp. 592–604.

[34] X.S.Zhou and S.T. Huang, “Image retrieval: feature primitives, feature representation, and relevance feedback”, IEEE workshop Content-based Access Image Video Libraries ,2000, pp. 10-13.

[35] F. Mokhtarian and S. Abbasi, “Shape similarity retrieval under affine transform,Pattern Recognition”, 35,2002, pp. 31-41.

[36] R. Xu and D. Wunsch, “Survey of clustering algorithms”, IEEE Transactions on Neural Networks”, 16 (3), 2005, pp. 645– 678

[37] J.M. Jolion, “Feature similarity, In Principles of Visual Information Retrieva”l, M.S.Lew, Ed. Springer-Verlog, 2001, pp. 122-162.

[38] B. H. Juang and L. R. Rabiner, **“**Hidden Markov Models for Speech Recognition”, Technometrics, 33 (3), 1991, pp. 251-272.

[39] J.T. Tou and R.C. Gonzalez, “Pattern Recognition Principles”, Addison-Wesley Publishing Company, Inc., 1974.

[40] G. Qian, S. Sural, Y. Gu and S. Pramanik, **“**Similarity between Euclidean and cosine angle distance for nearest neighbor queries” , SAC’04,ACM, 2004

[41] G.Salton, “Automatic Text Processing”, Addison-Wesley, reading,1989.

[42] X.S.Zhu and T.S.Huang, “Optimizing learning in image retrieval : a comprehensive review”, Multimedia System, 8 (6), 2003, pp. 536-544.

[43] F.Jing, M.Li, L.Zhang, H.J.Zhang and B. Zhang,” Learning in region based image retrieval”, Proceeding of International Conference of image and Video Retrieval (CIVR2003), 2003, pp. 206-215 .

[44] Y.Rui, T.S. Huang, M. Ortega and S. Mehrotra, “Relevance feedback: a power tool for interactive content-based image retrieval”, IEEE Trans. Circuits Video Technology, 8 (5), 1998,pp. 644-655.

[45] G.D.Guo, A.K. Jain, W.Y. Ma and H.J. Zhang, “Learning similarity measure for natural image retrieval with relevance feedback”, IEEE Trans. Neural Networks,13 (4), 2002, pp. 811-820.

[46] Y.Rui, T.S. Huang and S. Mehrotra, “Content-based image retrieval with relevance feedback in Mars”, Proceeding of the IEEE International Conference on Image Processing,1997, pp. 815-818.

[47] Krishna K. Pandey, N. Mishra, H. Sharma, “Enhanced of colour matching algorithm for image retrieval”, *IJCSI International Journal of Computer Science Issues*, Vol. 8, Issue 3, No. 2, 2011, pp. 529-532.

[48] L.Zhang, F. Liu and B. Zhang, “Support Vector Machine Learning for Image Retrieval”, International Conference on Image Processing, 2001, pp. 7-10.

[49] Y.Lu, C.Hu, X.Zhu, H.Zhang and Q.Yang, “A unified framework for semantics and feature based relevance feedback in image retrieval systems”, ACM International Conference on Multimedia,2000, pp. 31-37.

[50] P. Ekman and H. Oster, “Facial expressions of emotion”, Annual Review of Psychology, 30(1),,1979, pp. 527-554.

[51] M. Pantic and L. Rothkrantz, “Expert system for automatic analysis of facial expression”, Image and Vision Computing Journal, 18(11), 2000, pp. 881-905.

[52] M. Pantic and L. J. M. Rothkrantz, “Toward an affect-sensitive multimodal human-computer interaction”, In Proceedings of the IEEE, 2003, pp. 1370-1390.

[53] D. W. Oard and J. Kim, “Modeling information content using observable behavior”, 2001.

[54] J. Teevan, S. T. Dumais and E. Horvitz, “Personalizing search via automated analysis of interests and activities”, In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 449-456.

[55] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers and T. S. Huang, “Authentic facial expression analysis”, Image Vision Computing 25 (12) ,2007, pp. 1856-1863.

[56] S. W. Foo and L. Dong, “A boosted multi-HMM classifier for recognition of visual speech elements”, In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP ’03), vol. 2, 2003, pp. 285–288.

[57] J. J.Williams and A. K. Katsaggelos, “An HMM-based speechto-video synthesizer”, IEEE Trans. Neural Networks, 13 (4) ,2002, pp. 900–915.

[58] S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition”, IEEE Trans. Multimedia, vol.2, 2000, pp. 141–151.

[59] E. H. Adelson and J. R. Bergen, “Spatio temporal energy models for the perception of motion”, Journal of Optical Society of America, A 2(2),1985, pp. 284- 299.

[60] N. Petkov and E. Subramanian,” Motion detection, noise reduction, texture suppression,and contour enhancement by spatiotemporal Gabor filters with surround inhibition”, Biological Cybernetics, 97 (5-6) ,2007, pp. 423-439.

[61] J. Rocchio, “ Relevance feedback in information retrieval”, In: Salton G.Ed., The Smart Retrieval System—Experiment in Automatic Document Processing, Prentice-Hall, Englewood Cliffs, NJ, pp. 313-323.

[62] Y.Ishikawa and R. Subramanya, “MindReader: Query database through multiple examples”, in Proc. Of the 24th VLDB conference, (New York),1998.

[63] http://www.open-video.org/details.php

[64] http://Vireo.CS.CityU.edu.hk/VireoWeb81/

Biographies

About the Author---Mr. KARM VEER SINGH received his B.E. degree in Computer Engineering from Madan Mohan Malviya Engineering College, Gorakhpur, India in 1990, and the M.S. degree in Software Systems from Birla Institute of Technology (BITS), Pilani, Rajasthan, India in 2003. He worked as programmer in VBS Purvanchal University, Jaunpur, India from 1999 to 2011. He has joined at the post of System Engineer in Computer Centre, Banaras Hindu University in 2011. He is currently persuing Ph.D. in Computer Engineering from the Department of Computer Engineering, Indian Institute of Technology, Banaras Hindu University,Varanasi, India. His current interests are in the area of Pattern Recognition, Content-based Multimedia Information retrieval, Search Engines, Artificial Intelligence, Statistical Learning and parallel/distributed computing.

About the Author---Dr. ANIL K. TRIPATHI received his M.Sc. Engineering (Computers) from Oddesa National Polytech University, Ukraine in 1984, and Ph.D. in Computer Engineering from Institute of Technology, Banaras Hindu University, Varanasi, India in1992. Dr. Tripathi a Professor in Computer Engineering Department, Indian Institute of Technology, Banaras Hindu University, Varanasi, India has been engaged in teaching and research for last 26 years in areas of Software Engineering and Parallel/Distributing computing. One research monograph on “Scheduling in Distributed Computing Systems-Analysis, Design and Models” has been published by Springer USA in 2009. Two book chapters have been published from John Wiley (USA) and Springer (USA). He has more than fifty research papers in Journals and conference proceedings. Twelve scholars have been awarded PhD degrees under his supervision in fields of software engineering and parallel/distributed computing.