# Emotion based Contextual Semantic Relevance Feedback in Multimedia Information Retrieval

Karm Veer Singh
Department of Computer Engineering, Indian Institute of Technology, Banaras Hindu University,Varanasi, 221005, India

Anil K. Tripathi
Department of Computer Engineering, Indian Institute of Technology, Banaras Hindu University,Varanasi, 221005, India

## ABSTRACT
Every query issued by a user to find some relevant information,contains the semantic and its associated contexts, but, indentifying and conveying these semantic and context (present in the query) to MIR system is a major challenge and still needs to be tackled effectively. Thus, exploiting the plausibility of context associated with semantic concept for the purpose of enhancement in retrieval of the possible relevant information, we propose Emotion Based Contextual Semantic Relevance Feedback(ECSRF) to learn, refine, discriminate and identify the current context present in a query. We will further investigate: (1) whether multimedia attributes(audio, speech along with visual) can be purposefully used to work out a current context of user's query and will be useful inreduction of search space and retrieval time; (2) whether increasing the Affective features (spoken emotional word(s)with facial expression(s)) in identifying, discriminating emotions would increase the overall retrieval performance in terms of Precision, Recall and retrieval time;(3)whether increasing the discriminating power of classifier algorithm in query perfection would increase the search accuracy with less retrieval time.We introduce an Emotion Recognition Unit(ERU) that comprises of a customized 3D spatiotemporal Gabor filter to capture spontaneous facial expression, and emotional word recognition system (combination of phonemes and visemes) to recognize the spoken emotional words. Integration of classifier algorithms GMM, SVM and CQPSB are compared in ECSRF framework to study the effect of increasing the discriminating power of classifier on retrieval performance.Observations suggest that prediction of contextual semantic relevance is feasible, and ECSRF model can benefit from incorporating such increased *affective features* and classifier to increase a MIR system's retrieval efficiency and contextual perceptions.

## Keywords
Contextual semanticRelevance feedback, Spoken emotional words, Affective feedback, Facial expression, Multimedia Information Retrieval.

## 1. INTRODUCTION
Semantic and sensory gaps can be tackled possibly well by including the human user in the Relevance Feedback loop of the IR system that links signals to symbols. Bridging the semantic-gap is a core challenge in multimedia information retrieval field [1]. Based on such feedback information, the system updates a distance metric among the low-level features to better reflect high-level conceptual semanticdistances and modify the parametric space, feature space, semantic space, orclassification space to identify the relevant and irrelevant information. Various implicit [2,3] and explicit [4] relevance feedbacks and combination of implicit relevance feedback with explicit relevance feedback [5] methods have been proposed in recent years to deal with and to remove ambiguity of user's information needs, but not sufficiently as may be desired.

### 1.1 Problem Formulation
High rejection rate of multimedia query processing results in a serious issue because of unsuccessful attempts of the MIR system in terms of exploration of the search space or in other words the issue is to navigate through the search space using algorithmic and other approaches so as to reject the unwanted search sub spaces within the MIR system rather than providing such output which would be finally rejected by the query maker as requirement of query maker varies depending upon certain context (need, current task in hand, temporal interest and preferences of the concerned query). Every query is put forth by a user as per his/her requirement perceived in his/her mind.The complete context of the requirement may not get expressed in the query posed by user, whereas given the result of query processing the user's mind may reject results because it does not match with the semantic concept perceived in user's mind.The term semantic concept of some information is the interpretation of low level features(intensity of pixels, colour, etc) captured during the image filtering process into high-level concepts understandable by the user. Multimedia information has multiple semantic, depending upon context (need, current task in hand, temporal interest and preferences of the concerned query) hence we can attempt to exploit the plausibility of context associated with a semantic concept. Semantic concept, without context 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 matchplayed. 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 inwhich user wants to retrieve the information (footballmatch) 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 because of reduced search space. Thus, exploiting the plausibility of considering a context associated with semantic may be beneficial in retrieving relevance information. In [6], authors discuss challenges in a contextual retrieval approach. The context of the requirement may be exceedingly difficult to ascertain, identify, isolate and specify. Though, a query issued by user to find some relevant information, contains the semantic and its associated contexts, but, some major challenge arises and needs to be tackled when we try to identify and convey these semantic and context (present in the query) to MIR system. These are about what is the context? What would be a Context Descriptor? How it will be represented? How MIR system will capture, register and interpret up to some extents user's current context in the query? If the MIR system can capture, register and interpret up to some extents our reactions then it can possibly offer context-sensitive support for our interaction. To cope with the above discussed major challenges we

argue that audio, speech present in a query can be used to determine a user's current context to retrieve some information and can be used to increase the system's contextual semantic understanding of the visual contents. To characterize a context, audio and speech events that are highly relevant to specific semantic concept are collected and modeled. The context, we refer to in this work, would consist of the Audio, Speech and Visual present in the query clip and are of user's temporal interest and preferences. Mainly three types of contexts are possible one of them is *audio, speech context* and second is *visual context* and third is *joint context. Audio and speech context*is being named here as the *Primary Context* in this work,can be derived from the audio and speech features present in the query clip and its relationship to visual features set present in the Context Seed Training Database to be explained later in section 3.1. We further define a *Primary Contextual Semantic Descriptor* that represents context co-occurrence relationship between the *Primary Context* and visual features set representing a concept defined in Context Seed Training Database. A *Visual context*is being named here as the *Secondary Context*,can be derived from the visual features present in the query clip and its relationship to visual features set present in the Context Seed Training Database. We further define *Secondary Contextual Semantic Descriptor* that represents context co-occurrence relationship between the *Secondary Context* and visual features set representing a concept defined in Context Seed Training Database. The combination of these *Primary Context* and *Secondary Context* results in to the *Audio-visual context* that is being named here as a *Joint Context.*

Mathematically, we define set of primary contexts, $C^p = \{c_i^p : c_i^p$ is thei$^{th}$ primary context,$1 \le i \le n_p\}$ , $n_p$is the cardinality of $C^p$i.e.$n_p = |C^p|$and a set of visual features, $V_F^p = \{v_f^{c_i^p} : v_f^{c_i^p}$ is the visual feature corresponding to$c_i^p\}$. Further, we define set of secondary contexts,$C^s = \{c_j^s : c_j^s$is the $j^{th}$ secondary context,$1 \le j \le n_s\}$, $n_s$is the cardinality of $C^s$i.e.$n_s = |C^s|$ and a set of visual features,$V_F^s - \{v_f^{c_j^s} : v_f^{c_j^s}$ is the visual feature corresponding to$c_j^s\}$.

A mapping from a collection of primary contexts to a set of visual features represents a *concept* and is defined as$f_1 : C^p \rightarrow V_F^p$whichfollows $f_1(c_i^p) = v_f^{c_i^p} \forall$ i .

Similarly, a *concept* for secondary context $f_2 : C^s \rightarrow V_F^s$ can also be defined, in which notations follow the same pattern as above replacing $p$ by $s$ and i by j, and $1 \le j \le n_s\}$.*Primary Contextual Semantic Descriptor* can be defined by a relationship < concept, attributes> $= < f_1, (c_i^p, v_f^{c_i^p}) >$.In the same way, relationship< concept, attributes > = $< f_2, (c_j^s, v_f^{c_j^s}) >$ canbe used for *Secondary Contextual Semantic Descriptor.*

Now, a C*oncept* for joint context $f_3 : C^{ps} \rightarrow V_F^{ps}$in our work is defined as,

$f_3(c_{(i,j)}^{ps}) = v_f^{c_{(i,j)}^{ps}} \forall (i,j) \in (n_p \times n_s)$,

where$c_{(i,j)}^{ps}$is a joint context (consisting of both i$^{th}$ primary and j$^{th}$ secondary contexts) and$v_f^{c_{(i,j)}^{ps}}$is a visual feature which is $f_3 -$image of Joint Context$c_{(i,j)}^{ps}$.We define *Joint Contextual Semantic Descriptor* that represents context co-occurrence relationship$< f_3, (c_{(i,j)}^{ps}, v_f^{c_{(i,j)}^{ps}}) >$ between the *Joint Context*$C^{ps}$and visual features set$V_F^{ps}$ representing a concept in the Context Seed Training Database. These contexts will be utilized in improving the retrieval performance. Some of the contexts $c_i^p$, $c_j^s$ and $c_{(i,j)}^{ps}$may be in the query clip due to different audio, speech and visual features present therein. All attributes (audio, speech, visual and their combination i.e.( $c_i^p, v_f^{c_i^p}$), $(c_j^s, v_f^{c_j^s})$ and $(c_{(i,j)}^{ps}, v_f^{c_{(i,j)}^{ps}})$ in the query clip are being used to work out the context for the purpose of reduction of search space and hence retrieval time.Our aim is to design a MIR system that can capture, register and interpret up to some extents a user's current context in the query. Problem is to identify proper current contexts which may be one or more among attributes ( $c_i^p, v_f^{c_i^p}$), $(c_j^s, v_f^{c_j^s})$ and $(c_{(i,j)}^{ps}, v_f^{c_{(i,j)}^{ps}})$in a user's query and iteratively train a MIR system to refine these contexts that will be used in the query perfection to achieve the better performance. We, further, argue that perfection of context in the query can be done more successfully if emotional responses of a user are used as an implicit source of evidence in relevance feedback cycle to obtain prioritized context attributes, for example, reaction of a user in terms of positively favoring one type of context attributes where as reacting negatively for some other context attributes. Basic challenges, here, arise as to which implicit features will be useful? How should they be combined, and what are appropriate system responses that can be used to increase the discriminating power of differentiating emotions to identify the current context in a user's search query? How it can capture, register and interpret up to some extents a user's current context in query? It may be difficult and seem impossible to define analytically a contextual semantic. We can have only guess and estimate the semantic by using emotional reactions of a user when s(he) has shown interest on multimedia contents during retrieval process. Semantic can be guessed and estimated by capturing and interpreting a stream of user's actions (browse and navigation, eye tracking, etc) during interaction with multimedia contents [7,5]. It is suggested that relevance feedback can be utilized for affective retrieval [8] but still require an efficient affective feedback system that can capture, register and interpret up to some extents spontaneous reactions of a user and can use it to identify user's current context amongst multiple contexts present in a user's search query. How such a system will be designed? How we could employ affective feedback as an implicit source of evidence to increase the MIR system's contextual semantic learning, refining and discriminating contexts that will be used in query perfection to achieve the better retrieval performance in terms of precision, recall, search space and retrieval time? To cope with the above discussed challenges we argue that incorporating spoken emotional words (an example words list is given in Table 1.) along with facial expressions that facilitate a likely natural and meaningful emotion would be more effective implicit source of evidence in CSRF cycle and could allow a search system to predict, with reasonable accuracy, the contextual semantic relevance of information without the help of explicit knowledge.

**Table 1. Spoken emotional words**

| | |
|---|---|
| *Positive emotion words* | Vow, Hurrah, Wah, Yes, Yeah, Oh yes, ok, Beautiful, Superb, Mind blowing, Weldon, Fantastic, Lovely, Amazing |
| *Negative emotion words* | No, Oh no, Never again, Oh shit, Nonsense |

We propose Emotion Based Contextual Semantic Relevance Feedback(ECSRF) to learn, refine, discriminate and identify the current context present in a query. We will further investigate: (1) whether multimedia attributes(audio, speech along with visual) can be purposefully used to work out a current context of user's query and will be useful in reduction of search space and retrieval time; (2) whether increasing the Affective features (spoken emotional word(s) with facial expression(s)) in identifying, discriminating emotions would increase the overall retrieval performance in terms of Precision, Recall and retrieval time;(3)whether increasing the discriminating power of classifier algorithm in query perfection

would increase the search accuracy with less retrieval time. We introduce an Emotion Recognition Unit(*ERU*) that comprises of a customized 3D spatiotemporal Gabor filter to capture spontaneous facial expression, and emotional word recognition system (combination of phonemes and visemes) to recognize the spoken emotional words. Integrationof classifier algorithm GMM, SVM andCQPSB discussed in Section 3.4 are compared in ECSRF framework to study the effect of increasing the discriminating power of classifier on retrieval performance.*ERU* will be used to guess and estimate spontaneous reaction of a user to capture, register and interpret up to some extents a context in the query. Contextual Query Perfection Scheme refines the context on the basis of relevance judgment taken by *ERU* via relevance feedback. These refined contexts were used in the query perfection to reduce the search space hence the retrieval time and to reduce the ambiguity in semantic.

In Section 2, we introduce some recent work related to this research. Section 3 provides the details of our proposed *ECSRF* framework. 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

The contextual relevance feedback approach has well recognized research challenges in information retrieval. In [9], author has presented ongoing research on the implementation of the contextual relevance feedback approach in web-based information retrieval. In [10], authors study affective, facial and behavior reaction towards the picture contents. In [11], the relevance feedback is applied to reflect the user's emotion inevery retrieval processes. In [12], the existing EBIR applied to image retrieval by using the genetic algorithm. In [13], the authors explore the role of affective feedback in designing multimedia search system. Facial expressions have been associated in the past with universally distinguished emotions, such as happiness, sadness, anger, fear, disgust, and surprise [14]. State of the art algorithms for emotion recognition usually use the speech signal and/or the facial expression [15]. Few efforts have been done to link emotions to content-based indexing and retrieval of multimedia [16]. In [17], the concept of speech-assisted facial expression analysis and synthesis is proposed and provides useful information for expression analysis.

## 3. OVERVIEW OF THE PROPOSED 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 ECSRF 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 in the form of relationship< concept, attributes > in the database hereby known 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 3.1.1. 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 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 3.2. 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.

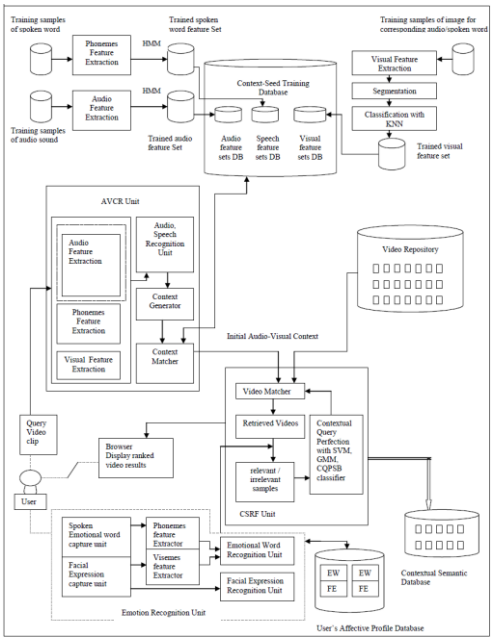


**Fig. 1: Proposed architecture of ECSRF framework**

## 3.1 Audio-Visual Context Recognition Unit

Extraction of rich features set will certainly increase the discriminating power of classifier and Affective feedback to identify the context in a query.Audio-Visual Context Recognition Unit is comprised of*Audio and Speech Recognizer*, *Visual Feature Set Extractor*, *Context-seed Training Database* and*Context Generator and Matcher*. The speech signal is realization of some message composed of basic sub-word lexical units, e.g., phonemes or syllables [18]. Whole word template are difficult to use for vocabulary recognition, because of whole vocabulary word must be spoken. So, in this propose paper, phonemes are chosen as sub-word unit. An example 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 Feature Set Extractor*, *Phoneme Feature Set Extractor*. The *Audio FeatureSetExtractor* pre-classify the speech and non-speech segment by using a KNN classifier and Linear Spectral Pairs-Vector Quantization (LSP-VQ) analysis [19] based on HZCRR ,LSTER, SF. From non-speech silence is detected based on STE, ZCR. Now*Phoneme Feature Set Extractor* detects speech boundaries from speech stream using algorithm[20] based on STE and ZCR features extracted from frame. The first 12 MFCC(static parameter) and 12 MFCC (dynamic parameter) are extracted from each frame to produce 12-D feature vector[21]. We adapt the phoneme-based vectorquantization technique given by [22] to generate the cluster (codeword). A Gaussian Mixture Model (GMM) is estimated for each phoneme by EM algorithm [23] that performsvector quantizationover the speech feature vectorand 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 [24] comprised of multiple N-state –M symbols HMM adapted as the sub classifier for discriminating Phonemes Feature Sets and used for word recognition system.

**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:/,/ |
| *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 |

Six texture features and six motion features are extracted by applying customized spatiotemporal Gabor filter given in Eq.(3), motion energy given in Eq.(5) and method discussed in Section 3.2.2. Nineteen edge features are extracted by applying edge filter and the Water-filling algorithm [25]. Eight shape features are extracted by applying algorithm discussed in [26]. 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 combination of these seven features sets makes the 70-dimensional low level feature vectors(texture,motion,edge,color,location,background-foreground features) and are being used by *Visual Feature Set Extractor*. Statistical clustering method like k-means or hierarchical clustering [27] is being used to classify the visual feature vectors in to number of clusters by mapping Euclidean / Gaussian visual similarity [28].

### *3.1.1 Context- SeedTraining Database*

Increasing the numbers of good training samples increases the discriminating power of classifier.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 that can represent the whole context knowledge are required but it is tedious, cumbersome yet impossible to collectsufficient and good training samples of audios, speech and visual objects and train the classifier to discriminate whole context accurately. We will study the effect of increasing the size of good training sample datasets and will try to find the answer of question "How many training sample sets of audio, speech and visual features would besufficient to recognize the whole context and fill the semantic gap?" in future work. However, we have developed initial *Context-seed Training Database* having trained AudioFeature sets DB,*Speech Feature Sets DB* and *Visual Feature Sets DB* that was used with Context Matcher in AVCRU to match audio-visual context. For this purpose, we have recorded sound of 4500 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 750 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 [29] 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 and stored in the Trained *Audio Feature Sets DB*. *Visual Feature Sets DB* is prepared by using *Visual Feature Set Extractor* and KNN classifier [30].Audio and speech recognized by *Audio Speech Recognizer Unit* provides the *Primary Context* and a visual objects recognized by *Visual Feature Set Extractor* provides the *Secondary Context*. Due to variations in audio and speech present in query clip, multiple contexts will be generated having multiple semantics of same visual. 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. 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. 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. We will study the retrieval performance in all the possible combinations of Audio-Visual contexts.

## 3.2Spontaneous Emotion Recognition Unit (ERU)

Our assumption is that, if a user shows emotions by speaking spoken emotional words and giving facial expressions during the information retrieval process, this can be the effective evidence to feedback cycle to predictthe reasonable accuracy of relevant and irrelevant information. In our companion work discussed in paper [31]multi modal approach (spoken emotional words and facial expressions) was used to model user's affective behaviour. In previous work [31], we have used spoken emotional words that use only Phoneme based words recognition system. In this proposed work, to increase the discriminating power of emotion we modify the ERU unit (a) with additional information from such as movement of lips (visemes: time sequence of lip movement)[34]). (b) with tracking and measuring any spontaneous deformation in

concave (local minima) and convex (local maxima) regions of face by customized 3D spatiotemporal Gabor filter and algorithm discussed in Section 3.2.2. Since a user is in RF loop and emotion of a user is subjective, we require a profile database of users to discriminate their emotions. We propose Emotional spoken word recognition System and facial Feature Extraction system to estimate and recognize the emotions and also to build *Affective Profile Database* of users.

### *3.2.1 Spontaneous Emotional Spoken Word Recognition System*

One objective is to design an emotional spoken word recognition system with small emotional words vocabulary consists of emotional words grouped in categories to express the six basic emotions and to evaluate the effectiveness of acoustic speech signals in detecting spontaneous changes in emotion during multimedia retrieval process. Audio-Visual integrated processing has been proposed for speech recognition [32,33]. The performance of ASR system can be significantly improved with additional information from visual element such as movement of lips (visemes), tongues and teeth. Like phonemes which are basic building blockof sound of language,visemes(time sequence of lip movement)[34] are basic building block for the visual representation of the words. Our proposed Emotional Spoken Word Recognition system uses the phonemes and visemes, a Gaussian like hyper ellipsoidal structure [22] treated as cluster in speech signal space, 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 and visemes, the system may be further trained to recognize words by breaking them in to sequence of phonemes and visemes. In *Phonemes Feature Sets extractor*, acoustic phonemes feature sets are extracted from audio signal by applying method discussed in Section 3.1. In *Visemes Feature Sets Extractor*, the*Visemes Feature Sets* is captured by motion energy, the similar method adapted in facial expression recognition discussed in later in Section 3.2.2. The extracted visual feature vectors are clustered by technique applied as in Phonemes Feature Sets Extraction in Section3.1 to generate a codebook entry for visem and forms a quantization code entry for a visem in *Visems Feature Sets*. *Phonemes Feature Sets* and *Visemes Feature Sets* are used for classification of emotional words. In this proposed recognition system, audio-visual classifier is implemented by late integration of two separate AdaBoost-HMM classifier [24] comprised of multiple N-state –M symbols HMM adapted as the sub classifier one for discriminating Phonemes Feature Sets and other for discriminating Visemes Feature Sets. Combination of these two, namely Composite AdaBoost-HMM classifier shown in Fig. 2, yields the final discriminating function and is used in training and recognizing the spoken emotional word.Several attempts have beenintroduced in thepast to build user's profileand learning user interest from implicit feedback[35,5].

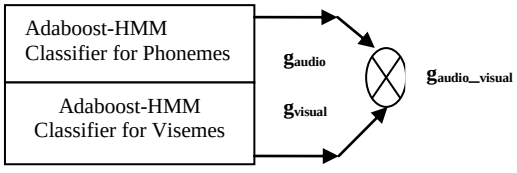


**Fig. 2: Composite AdaBoost-HMM classifier**

In our model, for contextual semantic relevance, we incorporate user's *Affective Profile Database* which stores the spoken emotionalwords and real facialexpression captured during the informationretrieval process. When recording emotional words and facial expressions several critical issues arise [36]. The most critical issue is that acoustic speech sound (phonetic), visual speech element (visemes) and facial expression may vary significantly from one individual to another (depending on personal, familial or cultural traits and on their vocal articulation). In our approach, taking into consideration the above factors we capture real facial expressions, sound of spoken emotional words and visems feature sets, thus increasing the chance of observing real spontaneous behaviour during retrieval process. Initially facial expression, utterance of emotional words classified as a certain phoneme or set of phonemes and corresponding visemes feature sets is stored in user's *Affective Profile Database*. During the retrieval process, actual facial expression, phonemes feature sets and corresponding visemes feature sets are extracted from the spoken emotional words. These facial expression, phonemes and visemes feature sets are actual representative training feature sets of six basic emotion categories for a particular user and refine the previously stored User's *Affective Profile Database.* Now this becomes the user's new emotional profile database.

### *3.2.2Computational Model for Spontaneous 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. The various methods to extract the features sets based on texture, shape, location and motion exist in literature [37]. 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 following proposed algorithm. The steps are as follows:

Step1: Facial portion of images separated from frames of video captured by webcam. Face matrix containing intensity values I(x,y) is obtained.

Step2: Apply Customized 3D spatiotemporal Gabor filter given in Eq.(3) designed and discussed inSection 3.2.2.1 on each pixel intensity denoted by I(x,y) of face matrix. Convolve it to get the response matrix of filter by applying Eq.(4).

Step3: (a) Find distance transform matrix of response matrix.(b) Find local maximas and local minimas of this distance transformmatrix. These local maxima and local minima pointsgive the center of convex and concave regions of face as discussed above. (c) Process each region's point matrix along with it's center. Start from center of one region and go to another adjacentregion by comparing intensity values (either increasing or decreasing) of all neighborhoodpixels in alldirections. Thisgives convex and concave region's points and boundary.

Step4: Find the spontaneous changes in position and direction of these convex and concave regions by applying motion energygiven by Eq.(5). These motion energy units obtained are matched with the motion energy units of real emotional trainingdatasets of users in their affective profile database, finally classified into basic emotion categories.

#### *3.2.2.1 Customized Spatiotemporal Gabor Filter*

A purely spatial filter considers information only at a single time instance and computes the output at a given time using only the input at that time. It cannot be used for motion analysis because motion is a spatiotemporal concept changes in space and time. In seminal work, [38] 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 [39] 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) in Appendix.Stimuli motion is three-dimensional motion. Stimuli can move in x, y, z direction simultaneously at particular time instance. To detect stimuli's motion in all directions at a particular time instance a filter of bigger receptive field is required. To get the bigger receptive field, we modify the spatiotemporal Gabor filter given in Eq.(1) in Appendix. Our modifications are two folds and as follows:

1. Center of Gaussian envelope moves withspeed Vc in all the three directions x, y, and z simultaneously.
2. Cosine wave travels with a phase speed V in all the three directions x, y, and z simultaneously.

Images in frames of video, captured by 2D video camera are two-dimensional in nature. The Intensity value of pixels increases or decreases gradually as the distance between stimuli and camera increases or decreases respectively. The difference in intensity value of neighbourhood pixels in a certain direction either increases or decreases as the relative distance between stimuli and camera decreases or increases. We consider the stimuli's movement towards camera or camera's movement towards stimuli as z-direction. We map the relative distances and motion of stimuli in z-direction by calculating the intensity difference between neighbourhood pixels in spatial domain (x and y direction). Although this mapping of relative distances in z-direction with intensity difference of neighbourhood pixels does not provide exact distance of camera to stimuli but with the help of this mapping we can predict the convex and concave region of stimuli. The mapping of relative distances in z-direction with the intensitydifference of neighbourhood pixels is denoted by Eq.(2) and shown in Fig. 3. (The Distance transform is applied on source image [40] downloaded[1] to show the convex and concave regions on face)

$$Z_d(x_i, y_i) = |I_A(x_i, y_i) - I_B(x_0, y_0)| 1 \le i \le n \quad (2)$$

where $I_B(x_0, y_0)$ is the initial value of maxima or minima as the case of convex point or concave point. $Z_d$ is the relative distance.We incorporate the effect of mapping of relative distances in z-direction with intensity differences of neighbourhood pixels to calculate the filter's response in z-direction as stimuli moves in z-direction.

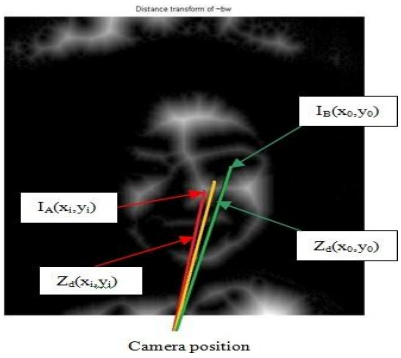


**Fig. 3: The mapping of relative distances in z-direction with the intensity difference of neighborhood pixels**

The modified spatiotemporal Gabor filter is given in Eq.(3) as follows:

$$g_{v,\theta,\varphi}(x, y, t) = \frac{\gamma}{2\pi\sigma^2} \exp\left(\frac{-\left(\sum_{x,y}(\bar{x} + v_c t)^2 + \gamma^2 \sum_{x,y} \bar{x}^2 + \gamma^2 Z_d^2\right)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda}\left(\sum_{x,y}(\bar{x} + vt) + Z_d\right) + \varphi\right) \times \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{(t-\mu_t)^2}{2\tau^2}\right) U(t) \quad (3)$$

Where

$$\bar{x} = x\cos\theta + y\sin\theta$$

$$\bar{y} = -x\sin\theta + y\cos\theta$$

$$Z_d(x_i, y_i) = |I_A(x_i, y_i) - I_B(x_0, y_0)|$$

$$U(t) = \begin{cases} 1 & \text{if } t \ge 0 \\ 0 & \text{if } t \le 0 \end{cases}$$

Parameters details are given in Appendix.

Here, the center of Gaussian envelop moves with Vc speed in all the three directions x, y, and z simultaneously. The cosine wave travels with a phase speed V in all the three directions x, y, and z simultaneously. The effect of movement of Gaussian envelop in z-direction and the effect of cosine wave traveling in z-direction is mapped with the intensity difference of neighborhood pixels in spatial domain. 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 I(x, y, t) is computed by convolution as follows:

$$r_{v,\theta,\varphi}(x,y,t) = I(x,y,t) * g_{v,\theta,\varphi}(x,y,t) \quad (4)$$

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 ($\varphi_2 - \varphi_1$) given below:

$$M_{v,\theta,\varphi}(x, y, t) = \sqrt{\{r^2_{v,\theta,\varphi_1}(x, y, t) + r^2_{v,\theta,\varphi_2}(x, y, t) + 2r_{v,\theta,\varphi_1}(x, y, t) r_{v,\theta,\varphi_2}(x, y, t)\cos(\varphi_2 - \varphi_1)\}} \quad (5)$$

## 3.3 Contextual Semantic Relevance Feedback

In RF a user judges the results as relevant and irrelevant to the query and machine learning algorithm is applied to learn the user's feedback then repeat the process until user satisfied with the results.The typical approaches in learning are 're-weighting' [41], query-point-movement QPM [41,42] and SVM is often used to capture a query concepts by separating the relevant images from irrelevant images using hyper-plane in projected space [41,43]. 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 semantic concept variability of user's perception in various contexts, the proposed RF system categorize the relevant and irrelevant contents by considering the contextual semantics. The proposed CSRF 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.

[1] http://www.kasrl.org/jaffe_download.html [Accessed 17.2.2012]

## 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 relevancefeedback. One of the most important and crucial step in CSRF is the third step that is machine learning algorithm to learn the contextual semantic andcan refine the context by taking user emotions as feedback.Thesystem dynamically learns the user's intention, and gradually present better search results. We have integratedvarious classifier algorithms GMM, SVM andCQPSB to evaluate and compare the effect of increasing discriminating power of classifiersin ECSRF framework on retrieval performance. 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 $l$, $l \le l \le \alpha$) in $PV^+$ having context $C_i$, and the set $A_{C_i V_l} = \{ V_l^{C_i} = V_l$, for some $l$, $l \le l \le \alpha \}$ with $\left|A_{C_i V_l}\right|$ = 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 $t$, $l \le t \le \beta$) in $NU^-$ having context $C_i$, and the set $B_{C_i U_t} = \{ U_t^{C_i} = U_t$, for some $t$, $l \le t \le \beta \}$ 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 algorithms as follows:

### *3.4.1 Contextual Query Perfection Scheme B(CQPSB)*

This scheme is based on Rocchio's formula[44] 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

$$Q_{new} = \alpha Q_{old} + \beta\left(\frac{1}{N_{R'}} \sum_{i=D_R} D_i\right) - \gamma\left(\frac{1}{N_{N'}} \sum_{i=D_N} D_i\right) \quad (6)$$

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 $\alpha, \beta, \gamma$ are selected constants. Our proposed Contextual Query Perfection Scheme B(CQPSB) 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

$$Q_{new}^{C_i} = \alpha Q_{old}^{C_i} + \beta\left(\frac{1}{N_{R'}^{C_i}} \sum_{i=D_R'^{C_i}} D_i^{C_i}\right) - \gamma\left(\frac{1}{N_{N'}^{C_i}} \sum_{i=D_N'^{C_i}} D_i^{C_i}\right) \quad (7)$$

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 $\alpha, \beta, \gamma$ are selected constants. In our proposed work $\alpha, \beta, \gamma$ all are having value 1. The values of $D'^{C_i}_R$ and $D'^{C_i}_N$ are as $D'^{C_i}_R = \left|A_{C_i V_l}\right|$ = n (say) that represents the set of +ve videos samples under context $C_i$ that is $V_1^{C_i}$ and $D'^{C_i}_N = \left|B_{C_i U_t}\right|$ = 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.

### *3.4.2Contextual Semantic Relevance Feedback based on GMM Classifier*

We consider a semantic concept as a class conditional probability density function over a feature space under a selected context. The true class conditional densities are not available, so assumptions must be made as to their type and parameters estimated during training data. A GMM[45] model is a type of probability density model which comprise of a Gaussian component function. A Mixture of M Gaussians having an n-dimensional observation vectors **x**is given as:

$$P(x|M) = \sum_{M=1}^{M} \partial_M \cdot \frac{1}{2\pi^{n/2} \left|\Sigma_M\right|^{1/2}} e^{-(1/2)(x-\mu_M)^T \Sigma_M^{-1} (x-\mu_M)} \quad (8)$$

Where $\mu_M$ is an n-dimensional vector, $\Sigma_M$ is an n x n matrix, $\partial_M$ is the mixing parameter satisfying $\sum_{M=1}^{M} \partial_M = 1$. To reduce the

classification time during relevance feedback, the estimation of Gaussian components and its parameters in GMM, we adopted the image retrieval method based on a set of coverings proposed in paper[46]. A covering in our scheme is a hyper sphere which contains as many positive samples as possible away from negative examples. The idea is to use the number of coverings as the estimate of the number of Gaussian components.let $V_l^{C_i}=\{V_{l1}^{c_i},...,V_{ld}^{c_i}\}$ that represents set of positive videos $(l=1,...,\alpha)$ in PV$^+$under context $c_i$.And let $U_t^{C_i}=\{U_{t1}^{c_i},...,U_{td}^{c_i}\}$ that represents set of negative videos $(l=1,...,\alpha)$ in NU$^-$under context $c_i$. The aim is to find a set of coverings

$s^{ci}=\{(s_m^{ci}, R_m^{ci}), m=1,...,M\}$ where, $S_m^{ci}$ and $R_m^{ci}$ is the center and radius of the m-th coverings. The maximum distance $d_{max}$ between $V_{l1}^{ci}$ and $V_l^{C_i}$, the minimal distance $d_{min}$ between $U_{t1}^{ci}$ and $U_t^{C_i}$ are calculated as: $d_{max}=\max_j D(v_{l1}^{ci}, v_{lj}^{ci}), \forall v_{lj}^{ci}\in V_l^{C_i}$ (9)

$$d_{min}=\min_j D(v_{l1}^{ci}, U_{tj}^{ci}), \forall U_{tj}^{ci}\in U_t^{ci} \quad (10)$$

The radius of coverings can be calculated as

$$r=(d_{max}+d_{min})/2, \text{ if } d_{min}\geq d_{max}$$

and $r=\rho.d_{min}$, $0\leq\rho\leq 1$ otherwise. The mean vector of each Gaussian component is as

$$\mu_m=\frac{1}{c_{V_l}^{ci}}\sum_{i=1}^{V_l^{ci}} x_i^m, \text{ m=1,...,M} \quad (11)$$

We assume that all dimensions of feature vectors are independent to each other so that covariance matrix $\Sigma_m$ is simplified to a diagonal matrix. Estimation of the covariance matrix is as follows:

$$\sigma_{jj}^m=\sqrt{\frac{1}{c_{V_l}}\sum_{i\in s^{ci}}(x_{ij}-\mu_j)^2} \quad (12)$$

### *3.4.3Contextual Semantic Relevance Feedback based on SVM Classifier*

SVM is often used to capture the query concept by separating the relevant samples from irrelevant samples using a hyper-plane in a projected space
[43]. Use of more discriminate learning approach requiring fewer parameters may yield better results for our experiment. We adopted SVMs with radial basis function kernels [47] for testing the effect of more discriminating power of classifier on retrieval performance under a context.

In our approach given training data
$\{(x_i,y_i)|x_i\in V_l^{C_i}, y_i=1; x_i\in U_t^{C_i}, y_i=-1; V_l^C=V_l^{C_i}\cup U_t^{C_i}\}$(13)
$V_l^{C_i}(U_t^{C_i})$ isthe positive ( negative) data set. It is to find a decision function that can separate the positive and negative pattern correctly. The programming of SVMs is as follows:
The best hyperplane can be found by minimizing

$$|w|^2+C(\Sigma_i\varepsilon_i), \text{ subject to}$$

$$y_i((w.x_i)+b)\geq 1-\varepsilon_i, \varepsilon_i\geq 0 \quad (14)$$

Where, C is a constant that controls the misclassification cost.It can be shown by using Langrange multiplier techniques that this is equivalent to minimizing a dual problem

$$\Sigma_i\alpha_i-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j\phi(x_i).\phi(x_j) \quad (15)$$

Subject to $\sum_i\alpha_i y_i=0, 0\leq\alpha_i\leq C$.

## 4. EXPERIMENTAL SETUP, RESULTS AND DISCUSSION

The development dataset for video repository contains a large number of videos provided by open-video project[2]. The video collection includes TRECVID 2001 and TRECVID 2002 videos datasets. 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[3]. Silent category videos were also downloaded from open-video project website to study the effect of occurrence of *Secondary Context* only 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. Nearly 500 hundred music video clips were also collected from Internet. After the collection of Video repository, user's *Affective Profile Database* and *Contextual Seed Training Database* a complete experimental setup and algorithm was implemented. We evaluated our ECSRF framework introduced in section 3. The system categorizes the retrieved video result sets into positive and negative sample sets based on positive and negative emotion recognized and discriminated 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. We followed each relevance feedback session for 5 iterations and measured the precision, recall and average retrieval time.

### 4.1 Study 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 MIR system by exhaustive retrieval testing. In our companion work discussed in paper [31], the framework was tested first without considering context, secondly considering *Primary Context* (audio and speech) only, thirdly considering *Secondary Context* (visual) only and finally considering joint context (audio + visual). We built precision vs. iteration diagram for each test. The precision of our proposed scheme, which combines audio, speech and visual as joint context, is shown in Fig. 4. Few observations were easily found.The combination of the audio, speech with visual as joint context consistently improved the understanding of contextual semantics. In specific, it improvedthe precision from 58% to 82% for our proposed scheme that uses *Joint Context* in comparison to *Secondary Context* (visual) only.We saw that joint use of audio and visual context signature provided significant improvement of contextual semantic concepts as compared to the use of visual

[2] http://www.open-video.org/details.php [Accessed 10.3.2012]

[3] http://Vireo.CS.CityU.edu.hk/VireoWeb81/ [Accessed 15.3.2012]

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. 5. In specific, the shortest vs. longest retrieval time was 0.52 minutes against 1.68 minutes. However, the corresponding precision was 82% and 65%.

## 4.2 Study the Effect of Addition of Affective Features in Discriminating Emotions in Contextual Semantics Relevance Feedback

We have argued that addition of Affective features in identifying, discriminating emotion will increasethe retrieval performance in terms of precision, recall and retrieval time. To validate this argument

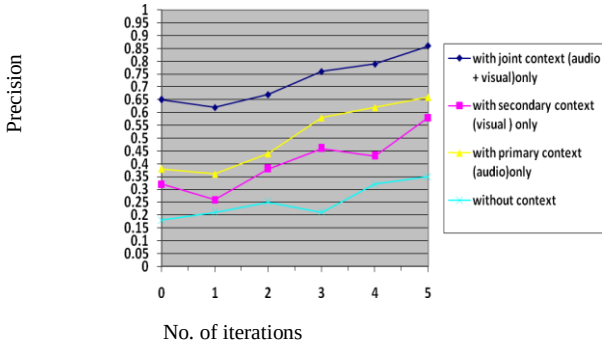


**Fig. 4:Precision comparison on considering occurrence of primary context only, secondary context only, joint context only and without context**

our proposed ECSRF framework in this work is tested first with CQPSB without addition of any Affective features, secondly, with CQPSB classifier with addition of visems in emotional word recognition as Affective features, thirdly, with CQPSB with additional spontaneous facial feature extraction (customized 3D spatiotemporal Gabor filter with motion energy) as Affectivefeatures and finally, considering joint Affective features(Addition of visems + Addition of Spontaneous facial feature extraction) with CQPSB.We studied the effect of Addition of Affective features in video information retrieval and contextual semanticunderstanding of MIR systemby

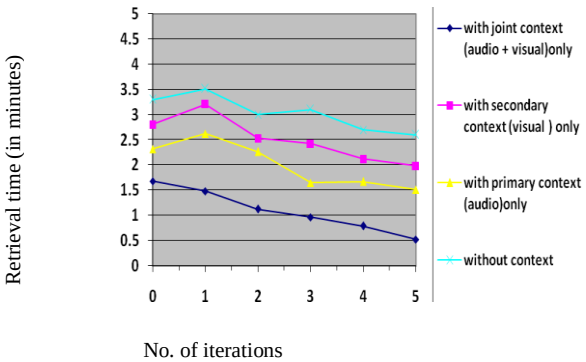


**Fig. 5: Retrieval time considering occurrence of primary context only, secondary context only, joint context only and without context**

exhaustive retrieval testing. We built precision vs. iteration diagram for each test andis shown in Fig. 6. These observations can be easily found.

(1) (a) The addition of visems in spoken word recognition system as affective features with CQPSB consistently improves the understanding of contextual semantics. In specific, it improves the precision from 66% to 72% for our proposed scheme that uses visems and phoneme feature sets as Affective features to discriminate emotional words as primary context.

1 (b) the addition of spontaneous facial expression extractor by customized spatiotemporal Gabor filter with CQPSB is also consistently improves the contextual semantic. In specific, it improves the precision from 58% to 82% for our proposed scheme that uses spontaneous feature extraction as Affective features in CSRF cycle as secondary context.

1 (c) the overall retrieval performance in terms of precision for our proposed scheme that uses joint context (spoken emotional word with addition of visems + spontaneous facial feature with customized spatiotemporal Gabor filter improves the precision from 35% to 86% in comparison to CQPSB without any affective features.

(2) The addition ofvisems and Spontaneous facial feature as Affective feature stabilizes the semantic fluctuation associated with CQPSB (without any addition of Affective features) based retrieval that are mainly due to less discriminating power to emotions, that is, the difference between the highest and lowest precision is 18% and 22% for our proposed scheme ECSRF with addition of Affective features as (*Joint Context* based) and *Secondary Context* (visual) based scheme respectively. We see that joint use of audio and visual context signature with addition of visems and spontaneous facial features provide significant improvement of contextualsemantic concepts in less retrieval timeas compared to framework in Section 4.1 with CQPSB(without addition of Affective features). This is mainly due to morediscriminating power of ECSRF(withadditional Affective features) than previous one. We also studied the effect of addition of Affective features 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. 7. In specific, the shortest vs. longest retrieval time is 0.36 minutes against 1.62 minutes. However, the corresponding precision is 86% and 68%.We found that use of joint context (spoken emotional word with addition of visems + spontaneous facial feature with customized spatiotemporal Gabor filter as Affective featureshad taken less retrieval time in giving the video result sets according to user's interest. This validates our proposed scheme and bestcompromises in terms of precision and efficiencyfor retrieval of video according to a user interestfrom large video repository.

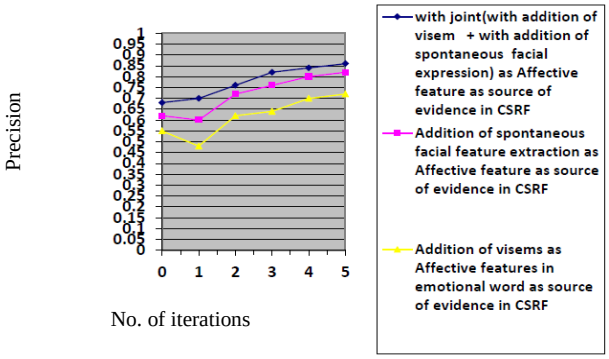


**Fig 6 : The effect of addition of visem in spoken emotional word and addition of spontaneous facial feature extraction as affective feature in facial expression as source of evidence in CSRF on Precision.**

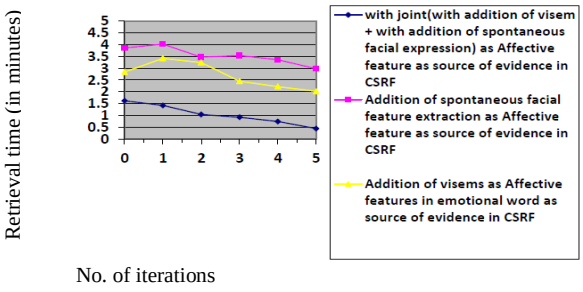


**Fig 7 :The effect of addition of visem in spoken emotional word and addition of spontaneous facial feature extraction as affective feature in facial expression as source of evidence in CSRF on Retrieval time.**

## 4.3 Study the Effect of Increasing the Discriminating Power of Classifiers in Contextual Semantic Relevance Feedback

We evaluated the effect of increasing the discriminating power of ECSRF framework by applying the classifiers GMM, SVM andcompared the retrieval performance in termsof precision,recall with CQPSB classifier discussed in a companion work[31]. For the purpose, we have taken 1500 videos as training set, 1000 videos as validation set and rest 2348 videos for testing set. We applied these classifiers algorithms for query perfection in ECSRF framework and to recognize the current context of a query. We have taken joint context (audio-visual) and ERU unit comprises of (spoken emotional words + spontaneous facial expression) to identify and discriminate the emotions that categorises the positive video samples and negative video samples. The refined context again used in next query. The precision vs. number of iterations diagram is shown in Fig. 8 for each method (ECSRF+GMM, ECSRF+ SVM and ECSRF+CQPSB) discussed above. In specific, ECSRF + SVM scheme provided the best precision 92.2% in compare to ECSRF+GMM(90.2%) and ECSRF+CQPSB(86.2%).This result comes mainly due to more discriminating power of SVM in compare to GMM and then CQPSB. We have also measured the retrieval time and built the retrieval time vs. number of iterations diagram that is shown in Fig 9. The shortest retrieval time for ECSRF+SVM,ECSRF+GMM and ECSRF+CQPSB are 0.36 minutes, 0.42 minutes and 0.44 minutes, respectively while longest retrieval time are 1.32 minutes, 1.46 minutes, 1.62 minutes. However, the corresponding precisions are 92.2%, 90.2% and 86.2%, respectively. We measured the accuracy, precision and recall of video retrieval for each case. The resultsare shown in Table 3.Thus, observations suggest that increasing the discriminating power of classifiers in contextual semantic perfection unit not only improve the system's contextual understanding hence increased precision but also reduces the retrieval time due to reduction in search space.

**Table 3. Considering joint context (spoken emotional word + facial expression) as source of evidences in RF**

| | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|
| ECSRF+ SVM | 92.6 | 92.2 | 88.4 |
| ECSRF+GMM | 90.4 | 90.2 | 86.6 |
| ECSRF+CQPSB | 86.6 | 86.2 | 80.4 |

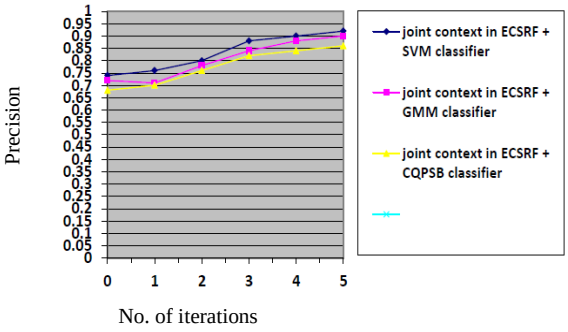


**Fig 8 : The effect of increasing the discriminating power of classifiers algorithm on Precision in CSRF.**

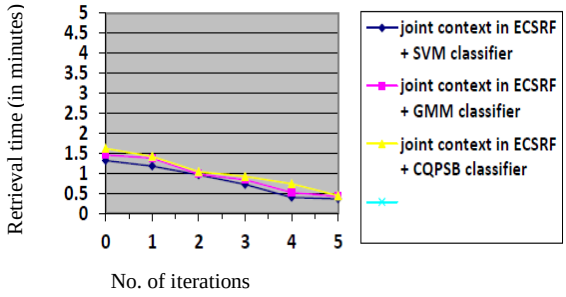


**Fig 9 : The effect of increasing the discriminating power of classifiers algorithm on Retrieval time in CSRF.**

## 5. CONCLUDING REMARKS AND FUTURE DIRECTIONS

In this paper, we have proposed Emotion Based Contextual Semantic Relevance Feedback(ECSRF) to learn, refine, discriminate and identify the current context present in a query. We further investigated (1) whether multimedia attributes (audio, speech along with visual) can be purposefully used to work out a current context of user's query and will be useful in reduction of search space and retrieval time; (2) whether increasing the Affective features (spoken emotional word(s) with facial expression(s)) in identifying, discriminating emotions would increase the overall retrieval performance in terms of Precision, Recall and retrieval time. (3) whether increasing the discriminating power of classifier algorithm in query perfection would increase the search accuracy with less retrieval time. We have introduced an Emotion Recognition Unit(ERU) that comprises of a customized 3D spatiotemporal Gabor filter to capture spontaneous facial expression, and emotional word recognition system (combination of phonemes and visemes) to recognize the spoken emotional words. We have integratedvarious classifier algorithms GMM, SVM andCQPSB to evaluate and compare the effect of increasing discriminating power of classifiersin ECSRF framework on retrieval performance. The result shows that this proposed emotion recognition technique has more discriminating characteristics in selecting relevant and irrelevant information and can be used to guess and estimate spontaneous reaction of a user to capture, register and interpret up to some extents a current context in the query. The feasibility of ECSRF framework was demonstrated for contextual semantic understanding. As shown by experimental evaluation performed on a large video repository, the use of joint Affective features(Addition of visems + Addition of Spontaneous facial feature extraction) with increased discriminating power of classifiers as source of evidence in our proposed ECSRF framework contributeda significant improvement in contextual semantic understanding of the MIR system and reducedambiguities in semantics.The test results support our hypotheses that multimedia attributes (audio and speech along with visual) can be

purposefully used to work out a current context for the specific query made by a user to reduce the search space hence the retrieval time and addition of Affective features (addition of visems in spoken emotional words with addition of spontaneous facial expression) and addition of discriminating power of classifiers could be effective evidence to RF cycle to increase the performance of retrieval in terms of precision, recall and retrieval time.The system is working fine but still requires a technique that can exploit the method discussed in this ECSRF framework and can be used to make a MIR system that can learn, refine, discriminate contexts in a query when multiple contexts co-occurring simultaneously in a query.Our future plan includes the work on scalability of the proposed scheme and its extensibility to tackle up to some extents user's subjectivity issue of contexts and can reduce the contextual semantic ambiguity under multiple contexts co-occurring simultaneously in a user's query.

## APPENDIX

The spatiotemporal Gabor filter is given in Eq.(1)

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

Where

$$\bar{x} = x\cos\theta + y\sin\theta$$

$$\bar{y} = -x\sin\theta + y\cos\theta$$

$$U(t)=\begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{if } t \leq 0 \end{cases}$$

$g_{v,\theta,\varphi}(x, y, t)$ 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 $\mu_t$ and standard deviation τ. Other parameters detail may be found in [39].

## 6. REFERENCES

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

[2] Hardoon DR, Taylor JS, Ajanki A, Aki KP, Kaski S (2007) Information retrieval by inferring implicit queries from eye movements. In Eleventh International Conference on Artificial Intelligence and Statistics

[3] Kelly D, Teevan J (2003) Implicit feedback for inferring user preference: a bibliography. SIGIR Forum 37(2): 18–28.

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

[5] Puolamaki K, Salojarvi J, Savia E, Simola J, Kaski S (2005) 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 ,pp 146-153

[6] Aytar Y, Orhan OB, Shah M (2007) Improving semantic concept detection and retrieval using contextual estimates. ICME

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

[8] Urban J, Jose J (2007) Evaluating a workspace's usefulness for image retrieval. Journal of Multimedia Systems 12(4-5) :355-373

[9] Limbu DK, Connor A, Pears R, MacDonellS (2006)Contextual Relevance Feedback in Web Information Retrieval. Information Interaction in Context, ACM, pp 138-143

[10] Lang PJ, Greenwald MK, Bradley MM, Hamm AO (1993) Looking at pictures: affective, facial, visceral, and behavioral reactions. Psychophysiology, 30 (3): 261-273

[11] Park JS, Eum KB, Shin KH, Lee JW(2003) Color Image Retrieval Using Emotional Adjectives. Korea Information Processing Society, B,10-B (2): 179-188

[12] Yoo HW, Cho SB (2004) Emotion-based Video Scene Retrieval using Interactive Genetic Algorithm. The Korean Institute of Information Scientists and Engineers, 10 (6): 514-528

[13] ArapakisI,Konstas I, Jose JM (2009) 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

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

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

[16] Salway A, Graham M (2003) Extracting informationabout emotions in films. In: Proceedings of ACMMultimedia '03

[17] Chang YJ, Heish CK, Hsu PW, ChenYC (2003) Speech-Assisted Facial Expression Analysis and Synthesis for Visual Conferencing System. Proceedings of ICME, pp 111 – 529

[18] Hayamizu S, Tanaka K,OhtaK (1988 A Large Vocabulary Word Recognition System Using rule based Network Representation of Acoustic Characteristic Variations. IEEE,1988

[19] Chang YJ, Heish CK, Hsu PW, Chen YC (2003) Speech Assisted Facial Expression Analysis and Synthesis for Virtual Conferencing Systems. IEEE, 2003

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

[21] Zheng F, Zhang G, Song Z(2001) Comparison of Different Implementations of MFCC. J. Computer Science & Technology, 16(6): 582–589

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

[23] McKenzie P, Alder M (1994) Initializing the EM algorithm for use in Gaussian mixture modeling. In Proc. Pattern Recognition, 1994

[24] FooSW, LianY,Dong L(2004)Recognition of Visual Speech Elements Using Adaptively Boosted Hidden Markov Models.IEEETransactions on Circuits and Systems for Video Technology, 14 (5): 693-705

[25] Zhou XS, Huang ST (2000) Image retrieval: feature primitives, feature representation, and relevance feedback. IEEE workshop Content-based Access Image Video Libraries ,pp 10-13

[26] Mokhtarian F, Abbasi S (2002) Shape similarity retrieval under affine transform. Pattern Recognition, 35: 31-41

[27] Xu R, Wunsch D (2005) Survey of clustering algorithms, IEEE Transactions on Neural Networks. 16 (3): 645– 678

[28] Jolion JM (2001) Feature similarity. In Principles of Visual Information Retrieval, M.S.Lew,Ed. Springer-Verlog,122-162

[29] Juang BH, Rabiner L R (1991) Hidden Markov Models for Speech Recognition.Technometrics, 33 (3): 251-272

[30] Tou JT, Gonzalez RC (1974) Pattern Recognition Principles. Addison-Wesley Publishing Company, Inc., 1974

[31] Singh KV and Tripathi AK (2012) Contextual Query Perfection by Affective Features Based Implicit Contextual Semantic Relevance Feedback in Multimedia Information Retrieval.IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3,pp. 191-202

[32] Morishima S, Ogata S, Murai K, Nakamura S (2002) Audio-visual speech translation with automatic lip synchronization and face tracking based on 3D head model. In Proc. IEEE Int. Conf. Acoustics, Speech,and Signal Processing, 2 : 2117–2120

[33] Silsbee PL, Bovik AC (1996) Computer lipreading for improved accuracy in automatic speech recognition. IEEE Trans. Speech Audio Processing, 4: 337–351

[34] Owens E, Blazek B (1985)Visemes observed by hearing impaired and normal hearing adult viewers. J. Speech Hear. Res., 28: 381–393

[35] Oard DW, Kim J (2001) Modeling information content using observable behavior. 2001

[36] Sebe N, Lew M S, Sun Y, Cohen I, Gevers T, Huang TS (2007) Authentic facial expression analysis. Image Vision Computing 25 (12): 1856-1863

[37] Bagherian E, Wirza R, Rahmat OK (2008)Facial feature extraction for face recognition: a review. IEEE,2008

[38] Adelson EH, Bergen JR (1985)Spatio temporal energy models for the perception of motion. Journal of Optical Society of America, A 2(2): 284- 299

[39] PetkovN, SubramanianE(2007)Motion detection, noise reduction, texture suppression,and contour enhancement by spatiotemporal Gabor filters with surround inhibition. Biological Cybernetics, 97 (5-6): 423-439

[40] Lyons M, Akamatsu J, Kamachi SM, Gyoba J (1998) Coding Facial Expressions with Gabor Wavelets. Proceedings, Third IEEE International Conference on Automatic Face and Gesture Recognition, IEEE Computer Society, pp200-205

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

[42] Guo GD, Jain AK, Ma WY, Zhang HJ (2002) Learning similarity measure for natural image retrieval with relevance feedback. IEEE Trans. Neural Networks, 13 (4) :811-820

[43] Zhang L, Liu F, Zhang B (2001) Support Vector Machine Learning for Image Retrieval. International Conference on Image Processing, 7-10

[44] 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.

[45] Bishop CM (1995) Neural Network for PatternRecognition .Oxford University Press, Oxford,UK.

[46] Zhang L, Lin FJ, and Zhang B (2001) A Neural network based self-learning algorithm of imageretrieval. Chinese Journal of Software, 12 (10):1479-1485

[47] Vapnik V (1995) The Nature of Statistical LearningTheory. Springer-Verlag, New York,NY, USA

## 7. AUTHORS PROFILE

**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. In March,2011, he joined Banaras Hindu University, BHU, Varanasi, India where he is currently System Engineer in Computer Centre. He iscurrentlypersuing his Ph.D. in Computer Engineering from 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.

**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 27 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. Fourteen scholars have been awarded PhD degrees under his supervision in fields of software engineering and parallel/distributed computing.