# An Approach to Model Pseudo-visual Retrieval System Based on Acoustic Signal for Visually Challenged People

**Pavan KUMAR**
**Department of Biotechnology, Indian Institute of Technology**
**Roorkee, Uttarakhand 247667, India**

**and**

**Durgesh KUMAR**
**MJK Retina Services, Sri Ram Janki Netralaya, Betiahata,**
**Gorakhpur, Uttar Pradesh 273001, India**

## ABSTRACT

This paper introduces the concept of modelling a visual retrieval system with certain requisite assumptions. A visually challenged person which has got the visual impairment due to some accidental cause or ailment at any stage of life has stored prior audio-visual coupled information about the daily life events in memory. The sole idea is to develop a retrieval system which utilizes the information collected from a visual device, moreover produces a peculiar audio sound which corresponds to visual data and capable of invoking the specific image into the brain which a person carries from early times. The consideration is to utilize the stored information pattern corresponding to an event into the same destination neuron in brain. A feasible concept of temporal distribution of audio bit pattern is taken into account to support the pseudo-visual retrieval system. An integrated approach of multi-staged information processing is required to develop a distinct acoustic signal, is considered to be of prime importance. Stereo-acoustic imaging approach is devised to reproduce the 3D frame of the image to some extent in terms of acoustic signal.

**Keywords**: Acoustic Signal, Pseudo-Visual, Information Pattern, Destination Neuron, Temporal Distribution and Stereo-Acoustic Imaging.
.

## 1. INTRODUCTION

The plan is to build up a strategy to invoke a pseudo-visual retrieval system to assist a visually impaired person who is just capable of hearing. Since light and sound waves are always two prime sensory dimensions to carry information to the brain of an individual, coupling of light and acoustic waves occur while sending the signal of an event to the brain. The two waves together then resulted into an input signal that goes into the brain following a specified path and finally stored into a destination neuron. Since both information patterns are to be stored at the same target a single key may be employed to retrieve the entire information of the event. The key of the sound wave information pattern of the event is sufficient to recall the sole event. The neuro-physiological tasks of visual conversion of sensory information signals from two dimensional to three dimensional representations in a semiotic way is performed in left occipital regions particularly L:17-18 of the brain. There occurs the role of Broca-Wernicke part in the conversion of auditory sensory signals to number of sound types thereby forming neurolinguistic sort of patterns. A class of ergodic distribution of visual semiotic language is used to occur within the brain that seems to have some alphabet of images. The complex pattern operation in two major modalities viz. electromagnetic resonating mode and morphological type mode are often resumed with a finite set of such alphabets of fundamental images. Any trail associated to recall of a visual form of memory implies a biological identification through which are actually existing in the form of quantum encoding in the biochemical aspects of microtubulin proteins present in neuropil. The present effort of computing the human sensation indeed entails the local and global binding networks of such semiotic patterns. Various probabilistic modelling of audio video data has been already done to develop an audio-video adaptive model comprising of Gaussian mixed approaches. Actually, visual data acquired by a fixed camera can be easily supported by audio information allowing a more complex analysis of the monitored scene.

## 2. AUDIO-VISUAL APPROACHES

Audio-video information is typically conveyed via the hepatic and auditory channels [1]. An electronic travel aid device can enable blind individuals to “see the world with their ears”. The visual aid device will provide a way to generate a correspondent sound providing the auditory cues for the perception of the position and distance of the pointed surface.

To increase the precision of tracking, a forecast of the position of the pointer in the image is used assuming a constant speed model. The most common choice is to take the closest candidate, as regarded by the Nearest Neighbour Data Association algorithm [2]. The sounds that give directional cues are said to be highly “spatialized”, in other words they provide distance, azimuth and elevation cues with respect to position of the object [3]. Such model adopts a versatile structural model for rendering of the azimuth [4] or orientation. The usability tests of such aid device are planned, with a group of both sighted and visually-impaired subjects. Precision and latency of the system will be measured systematically.

Now a day's Sounding Landscape method is of great use in processing acoustic signal on orientation based technique. Human computer interaction technology is simultaneously taken into account. Electronic Travel Aids (ETA) is used and precision and latency of the system are considered together with finite-distance scheme technique [5]. Waveguide mesh and waveguide filters [6] are applied for various screening purposes related to signal processing. The computation of coordinates of a 3-D point object nearest to two projected rays from a stereo camera can be done from the resulting triangulation [7]. Laser tracking and signal to noise ratio is computed to take any inference regarding error generation. Standard visual inspection systems are constructed to optimize and control the optical methodology. Probabilistic modelling of audio and video data has often been done to rectify the data set. Integrated approach of using coupled audio video adaptive model based on Adaptive Gaussian Mixture Model is being implemented to develop a multilevel probabilistic framework. Multimodal model modelling of scenes, working on-line is considered using one static camera and one microphone. Detection of single auditory visual event at a time in audio-visual background and foreground situation is often considered.

Sleeping foreground problem is very relevant and it can be observed as, foreground (the moving object) gets stops - integrated to background model. Computational Auditory Scene Analysis (CASA) and Computational Auditory Scene Recognition (CASR) are the two techniques which are generally used for this purpose. Multi band spectral analysis of the audio signal at video frame using Yule-Walker auto regressive method is often assumed. Multimodal data set, video frame, audio signal analysis and temporal frame shift application are the general assumptions to build up real time models [8]. Audio-visual matrix information together with BG - Background scene (static scene), FG - Foreground scene (objects acting in scene) and FG detection technique is applied to construct a spatial frame [9].

Three steps are applied for probabilistic modelling, firstly audio-visual time adaptive per-pixel mixture of Gaussian model, secondly foreground histogram model to classify foreground events and third one is adaptive mixture models operating on audio-visual data. Human Visual System (HVS) and retrieval systems like Content Based Image Retrieval (CBIR) are used in standard applications. Background and Foreground concept is designed to consider the discrete objects. Image processing and audio pattern processing are the different ways that have been frequently implemented to govern problems arising in selecting the proper range of wavelength.

Image retrieval system algorithm aims to discard irrelevant images and increase the amount of relevant one in the database. Two staged ant colony algorithm is the advanced methodology with two steps to find optimal path to food, secondly higher ranked images utilized for final retrieval. The synergy of low level descriptors is considering as group of ants seeking the optimal path to the food. Final retrieval is performed from the resulted aggregated deposition of pheromones. Effective method of matching the images is through the intersection of their colour histograms. Low level features widely used for indexing and retrieval of images. Spatially- biased histogram concept is used with global histograms for image retrieval proved to be an efficient and robust retrieval method. The whole image is convolved with the M- matrix resulting in new hue component which contain the colours information for the neighbourhood of each pixel.

Mostly researches pertaining to image retrieval system are utilizing Ant Colony Optimization (ACO) Heuristics [10] and Travelling Salesman Problem. In the prior approach the governing parameters such as optimum number of ants and pheromone decay rate have major relevance. Use of artificial ant colony as an optimization tool in the field of image retrieval is of great trend. The ants walk depends strictly on the information of relative position and distance of the surrounding food. Matusia distance concept is used considering the query histogram, the histogram to be compared and the number of bins. The pre-classification of thousands of image database is done and concluded to a pool of just few hundred, where the final retrieval takes place. The possibility of taking false positive result is also taken into account. The final retrieval is not simply based upon the value produced by the metric stating the distance between the features of the images but depends on the final amount of pheromone that has accumulated on each image considered in two different stages; viz. firstly, extraction of the descriptors from the image in the database and considering each histogram bin to be a virtual ant, secondly, the approach comprise of classification of image database for final retrieval. The database is there to measure the system's effectiveness and efficiency. One of the world largest database, LabelMe databases of images available freely on the internet because of immense volume which is adequate for various testing purposes. The image sets are the characteristic of the image database and also show the worst and best precision and recall the performances of the proposed system. Firstly the pooling of the images is done and principle followed that the highest ranked image is liable to acquire the most pheromone.

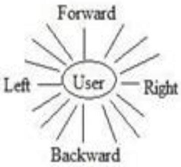
Basic elements of three image descriptors: a newly proposed spatially biased histogram and second histogram acquired through the utilization of the chromaticity-texture-energy of the image and lastly a histogram which results through the incorporation of certain attributes of the human visual system. Various methods are in use applying histogram, convolution and energy volume are kept in view to handle the audio-visual correlation technique in a more sophisticated way.

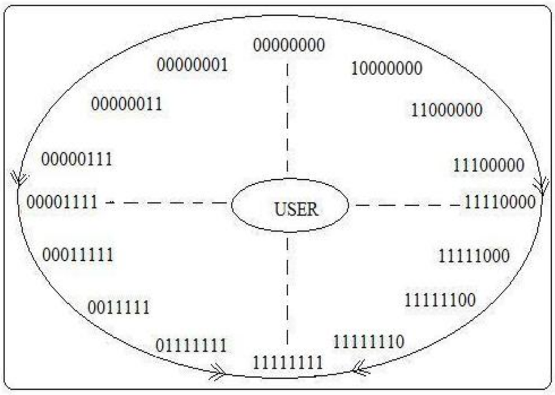
Automated surveillance systems have acquired an importance in recent years with multimodal model. An auditory vision substitution approach primarily aims at conveying the visual information utilizing the sense of hearing. These several approaches of vision technology are useful in sensory substitution and intend to provide a new approach of non-invasive visual prosthesis which is a sort of synthetic vision. Using this technique a live camera view is converted into sound scapes or a mixed type sound. In this system video to audio mapping is thereby correlating height to pitch of sound and brightness to loudness. A provision is there to refresh the views once per second using an approach of typical image resolution and spectrographic observation. . The present neuroscience research has proven that the adult blind people can made well responsive to sound with little training because of having a developed visual cortex, so an approach of “seeing and walking with pitch and loudness” might be enforced merely using such fixed devices.

Also efforts have been made recently to develop visual to auditory substitution device for prosthesis substituting vision for auditory mode. It uses a device that carryout real time translation of visual patterns into audio form. On the movement of user the device take visual frames at a high frequency and generates the corresponding complex sounds that allow recognition of the object. The captured visual stimuli are transformed into auditory stimuli with the use of a system that has pixel to pitch relationship and behaves like a rough model of the human retina while seems equivalent to an inverse model of the human cochlear apparatus. This software produces a typical sound which is a mixture of sinusoidal sounds produced by certain virtual sources against each of the receptive field in the image, a set of localized pixel give rise such receptive field.

## 3. NEW MODEL ASSUMPTIONS

A blind person has two major problems in initiating a movement step from the rest position, these are basically the static hurdles around him and to locate the direction from which a dynamic object is approaching, secondary problems are object dimension, approaching speed and surface topology of ground etc. In order to plan a strategy for the prior one, if the user has given an earphone which can carry the amplified natural sound conjugated with any or both of the (one for the presence of the static objects and another for the orientation of approaching objects) two characteristic sounds at a time, also there may be an OK sound tune as usual during the movement to justify his movement after testing the criteria of presence and approach of any hurdle, and one bell is just to stop any movement. Moreover the algorithm also comprises of testing the dimensions of objects that are fixed or approaching to support a permissible move. If we design a hypothetical frame for any particular time instant of a user it looks (shown in **figure 1** below) feasible that the direction of the approaching object can be evaluated simply in bits to signify the approach in a considered X-Y plane and individual bit string is assumed to correspond to a characteristic sound. If we consider a bit pattern of length eight signifying more uniformly for forward & backward as well as for left & right directions, now the temporal distribution in a horizontal frame are simply assumed to be equally divided and mentioned in a form of regular variant of bit patterns. Now the pitch of sound corresponding to this pattern distribution has a sequential pitch variation, so any user can be made familiar to the standard pith for just four directions and he may easily make decision to recognize the probable direction for other characteristic sound pitches belonging to this distribution, shown in **figure 1**. The noise in the signal due to presence of distant objects out of frame or tiny things within frame are to be removed using statistical approach on simulation. The frame is considered to be moving with the user and he receives information only when the relative velocity of approaching object is significant in respect to the user’s position at a particular time instant. Moreover an additional approach of identifying common object is to be put in use to solve the “indoor problems”. This is particularly associated to the day to day life movements for most of the problems which a person can face most of times. If a programmed camera is mounted in use which is capable of detecting the most frequent encountering objects like window and door etc. The recognition of such objects can be done taking into account the dimension and consistency based information analysis, with fixed information already incorporated in device. This will aid in promoting the general life activities where the forth most requirement is to recognize doors, beds etc situated in the room. For the same reason particular speaker could be devised to ensure the audio tune signifying the detection of the daily life usable things and indicating their presence in the room on coming to a certain approachable distance from the user.





Temporal distribution of bit patterns signifying the direction of approaching an object to the observer in an X-Y plane within a moving frame at a certain time extent **Figure 1.**

The view is to collect a requisite data for almost all possible general life events. The data-set is then used for further information processing. Training is required for the data set that has the recorded data for a bulk of possible usual events in our surrounding. An amplified and distinct sound can be produced in general audition pitch for the person or user to aid in recognition of the object in his surroundings. A radar based camera or detector (utilizing Doppler Effect and ultrasonic wave) is to be utilized to read the movement of objects in its surroundings. The user should be first trained for the specific sounds that are meant for different special orientation and position (moving or fixed) of different objects at a continuous short-time span. The audible signal then reaches to brain and the person will have temporal information of the all minor instants getting a view spatial-distribution of objects in all around.

### Moving frame assumption

Likewise characteristic pitch and loudness of the sound is to be devised to locate the direction and dimension of the static objects coming in the moving frame of observer. So on being trained to audio perception against a relative movement frame recording the real-time surrounding the user may either stop or deviate on getting close to a threshold distance from an object thereby listening the changes in pitch or loudness corresponding to the video frame. Also perception to spectral dimensions is to be imparted in the camera so that on colour basis the peculiar image can be distinguished and the user can be able to get a specific audio signal for the same. It would be helpful in identifying the grass lawn, road edge etc in outside environment while door, table like indoor objects. A simple algorithm is represented in a flow diagram below stating the entire layout of information processing, **figure 2.**

The information modalities easily available could be effectively used to generate complementary information in respect to visual data, in order to discover "activity pattern" in the background scene. Multimodal background model introduced together with video data and audio processing performing an auditory scene analysis.

BACKGROUND (STATIC) / FOREGROUND PICTURE (MOVING OBJECT) OBTAINED BY RADAR / INFRA RED BASED CAMERA

⇩

VISUAL DATABASE PROCESSING (IMAGE SCREENING)

⇩

AUDIO-PATTERN GENERATION IN 3D FRAME

⇩

AUDIO-PATTERN DATABASE PROCESSING (ACOUSTIC-SIGNAL SCREENING AND AMPLIFICATION)

⇩

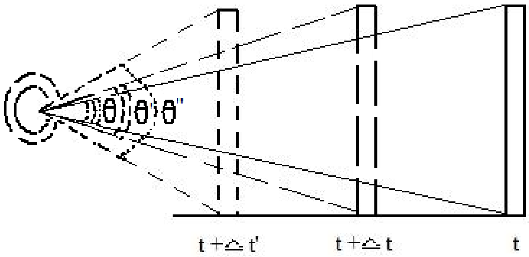
BACKGROUND SOUND + CHARACTERISTIC SOUND GENERATION

⇩

AUDIO SIGNAL THROUGH MICROPHONE

Audio-visual retrieval system design involving the different stages of information processing
**Figure 2.**

The on line synchronous audio-visual pattern is a method for integrating audio and video information. Various time-adaptive Gaussian methods together with visual background and visual foreground or moving object detection method are used for audio background modelling. The audio-visual fusion approach is applied to synchronize the two data. The relative change can be measured in terms of change in angle with respect to time (**figure 3**), hence generating information regarding the size, distance and speed of the object approaching (or moving away) present in the vicinity of the user.



Representation of 2D angular variance due to relative movement between object and mounted camera, present with the user.
**Figure 3.**

### Stereo-acoustic signal generation

The stereo-acoustic signal is generated which is utilized in performing the stereo-acoustic imaging of the obtained audio signal in respect to the scanned static (which serves as background) and dynamic objects as foreground which are scanned using Doppler's effect principle particularly relative movement of any object in the assumed frame is taken. The nature or the consistency i.e. hardness or softness of the present object in the scene can be well identified by the use of infra-red camera which can recognize the natural frequency emitted by the two types of objects present in the moving frame. The reading of colour, texture, shape and size etc can be helpful in identifying the basic items or their presence in surrounding. It is only possible with the use energy volume concept, it is generally governed with energy or acoustic emission based technique. There occurs energy transformation utilizing wavelength, frequency domain etc. All images are transformed into to a assumed space with *x', y'* etc components, each corresponds to a particular characteristic. Different vectors *i, j* and *k* are supposed to carry information regarding the energy component, **Eq. (1)**. A real time matrix based computation can be done to compute the energy volume signifying the orientation of surrounding. The designing of stereo frame of any closed vicinity at a particular time *t* depends upon the total energy volume (*ET*) computed from the energy of static and dynamic things (as *ES* and *ED* respectively) which are present in the assumed time dependent frame, given by E**q. (2**).

$$\mathrm{EN_V} = \sum \left\{ \sum_{i=1,j=1,k=1}^{p,q,r} \mathrm{E}_{x'}(\mathrm{i,j,k}) \;, \sum_{i=1,j=1,k=1}^{p,q,r} \mathrm{E}_{y'}(\mathrm{i,j,k}) \,, \ldots \right\} \qquad \text{..................} \quad (1)$$

$$ET(t) = ES(t)_i + ED(t)_j \qquad \text{....................} \; (2)$$

The requirement is there to synchronize the visual signal processing to audio signal generation/processing and finally the transmission of processed acoustic signal to the user by a microphone. A time dependent binary, 3-bit encoded hypercube concept is shown in **figure 4,** which incorporates the implementation of stereo-acoustic 3D imaging of any object present or coming into binary frame, represented with different time instants. It would be helpful in detecting and measuring the disturbances in a background scene via collecting information due to moving things in a defined frame.

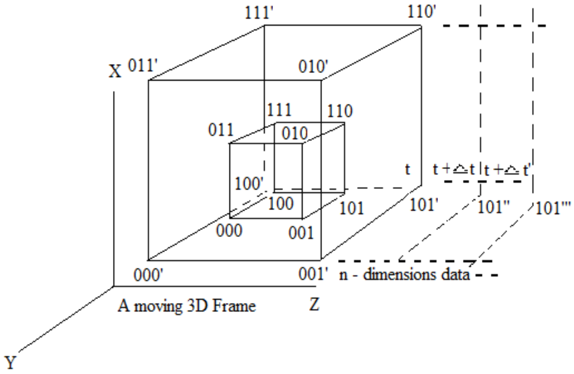


Diagram showing a moving stereo-acoustic 3-D binary frame which has been shown varying with time as stereo camera approaching the object (t➞ t +△ t, t +△ t'). **Figure 4.**

For processing our shortly collected information computational intelligence techniques like Artificial Neural Network can be used to bring the required accuracy. Thus every new sound produced by the device can invoke a visual sensation for the corresponding sound signal; the person can easily entertain and respond to the activities in his surrounding getting aware of the time dependent dynamic disturbances and the static barriers. Implementation of this idea requires a recording camera specially, a radar based and infra red based camera with multiple probes for different directions, a computing device or chip having stored data and software to process it; moreover an earphone is to be used to hear the audible signals. It looks to be a better approach than using sensor sticks that impose an extra-burden of carrying it all the time.

## 4. CONCLUSIONS

The people with fatal diseases like Retinal-Pigmentosa and Retinitis etc which often results into the loss of vision at any stage of life, may certainly get support to some extent through this proposed technology. Such patients have all previous experiences and corresponding information stored in some part of the brain. Now the acoustic information patterns together stored with visual patterns can be well utilized to make feasible this stereo-acoustic approach. It is evident that the birth-born blindness is less fatal in comparison to accidental or aliment-resulted blindness since an abrupt psychological deterioration occur which results into the impairment of sole mental and physical health. So any such approach may be of great worth to provide some assistance to such people.

## 4. REFERENCES

[1] R. Fish, **An audio display for the blind**, IEEE Trans. Biomed. Eng., 23(2), 1976.

[2] Y. Bar-Shalom and T. Fortmann, **Tracking and Data Association**, AP, 1988.

[3] D. R. Begault, **3D Sound for virtual reality and multimedia**, AP Professional, Massachusetts. Avenue, Cambridge, 1994, pp. 955

[4] C.P. Brown and R. O. Duda, **A structural model for binaural sound synthesis**, 6(5) September 1998, pp. 476- 488.

[5] F. Fontana, A. Fusiell, M. Gobbi, V. Murino, D. Rocchesso, L. Sartor and A. Panuccio, **A Cross-modal Electronic travel Aid Device**, LNCS 2411, 2007, pp. 393-397.

[6] A. Scot, V. Duyne and J. O. Smith, **Physical modelling with the 2-D digital waveguide mesh**, Tokyo, Japan, 1993. ICMA, pp. 40-47.

[7] E. Trucco and A. Verri, **Introductory Techniques for 3-D Computer Vision**,

Prentice-Hall, 1998.

[8] M. Cristani, M. Bicego and V. Murino, **Audio-Video Integration for Background modelling**, ECCV 2004, LNCS 3022, 2004, pp. 202-213.

[9] M. Kubovy and D. Van Valkenburg**, Auditory and visual objects**, Cognition, 80, 2001, pp. 97-126.

[10] K. Konstantinidis, G. C. Sirakoulis, and I. Andreadis, **An intelligent Image Retrieval System Based on the Synergy of color and Artificial Ant Colonies**, SCIA 2007, LNCS 4022, 2007, pp. 868-877.