Biomedical & Pharmacology Journal Vol. **10**(1), 247-254 (2017)

Visually Impaired Voting Aids Using Speech Processing & Face Recognition

A.KARTHIKEYAN, M.MAHENDRAN, S.A.K.JAINULABUDEEN and K.SOMASUNDARAM

Department of CSE, Panimalar Engineering College, Chennai, India.

http://dx.doi.org/10.13005/bpj/1104

(Received: December 27, 2016; accepted: January 16, 2017)

ABSTRACT

In proposed method, first voice is given as the input (The voice will be the adhar card number) and converted in to text in order to compare with the database. If the adhar card number matches, then the face recognition will be taken and compared. If both adhar card number and face will be authenticated, then the already recorded voice will be played. The recorded voice will have the details about the political parties, and according to the recorded voice we want to give the input. Then the vote will be done accordingly. In this system HMM, GMM and SOM (self organized mapping) filters are used for speech. Segmentation analyzes the system MFCC with SVM trained samples of speech and recognized text rate results given via MATLAB IDE. Efficient results give which compare to existing HMM model. For face identification using viola Jones algorithm is used.

Keywords: Mel-Frequency cepstral coefficients (MFCC), Hidden Markov Model (HMM), Gausian Mixture Model (GMM), Acoustic vectors and models, Dynamic Time Warping.

INTRODUCTION

The visually impaired voting aids are helpful for blind people in casting their vote. In current voting system Electronic Voting Machines are used in Indian state election to cast the vote the blind people votes are being casted without there knowledge. To overcome this problem a new voting aids is developed using speech processing and face recognition this eliminate the casting of wrong vote the person speech is used to cast their vote without any help the voters details are stored in the database and they are accessed at the time of voting face recognition is used to recognize the persons face and match with the image in the voter card. Speech processing is the study of speech signals and the processing these signals. The signals are processed in a digital form. Speech processing is also called as digital signal processing that can be applied to speech signal. The voter's voice is recognised using speech recognition. It analyzes the person's specific voice and uses it to tune the recognition of the person's speech, which results in increased accuracy. Systems require training before speech is recognised those systems are speaker dependent some systems does not require training they are speaker independent speech recognition is used in various applications like voice dialling call routing simple data entry speech-to-text processing accuracy of speech recognition varies in various environment if the vocabulary size is increased then the error rate increases A speaker-dependent system is used by a single speaker. A speaker-independent system used by any speaker. Various types of speech are available like isolated speech continuous speech discontinuous speech in isolated speech single words is used so it is easier to recognize the speech. So it becomes easier to recognize the speech like isolated speech. In continuous speech naturally spoken sentences are used, so it becomes harder to recognize the speech.

Various algorithms and techniques are used to increase the accuracy of the speech. Hidden Markov models (HMMs) are used widely in many systems. Language modelling is also used in many natural language processing applications. Modern g speech systems are based on the Hidden markov Models. HMM are statistical models that produce output as a sequence of symbols or quantities. HMMs are popular because they can be trained automatically and computationally feasible to use. Markov model would output a sequence of *n*-dimensional real-valued vectors (with *n* being a small integer, such as 10), outputting one of these every 10 milliseconds. A large-vocabulary system would require context dependency for the phonemes. The cepstral normalization is used to normalize for different speaker and recording conditions; for further speaker normalization it uses vocal tract length normalization (VTLN) for male-female normalization and maximum like hood linear regression (MLLR). Decoding of the speech would probably use the viterbi algorithm to find the best path, there is a choice between dynamically creating a combination hidden Markov model, which includes both the acoustic and language model information, and combining it statically .another approach is Dynamic time warping (DTW)-based speech recognition In this approach for measuring similarity between two sequences that may vary in time or speed. For instance, similarities in patterns can be detected, even in one video if the person was walking slowly and if in another they were walking more quickly, or if there were accelerations and deceleration during the course of one observation. DTW has been applied to video, audio, any data that can be turned into a linear representation can be analyzed with it. Modern speech recognition systems use e Mel-frequency cepstral coefficients or per-ceptual linear prediction as their aqustic features .Novel features are developed based on estimating a Gaussian mixture model (GMM) from the speech spectrum. The parameters of GMM are estimated using the EM algorithm. Various forms of GMM feature extraction are extracted, including various methods to enforce temporal smoothing and to incorporate a prior distribution to constrain the parameters. There are two noise compensation methods one during the front-end extraction stage and the other a model compensation approach. Face recognition is used to detect the human face and match with the data base. Some algorithms identify facial features by extracting landmarks, features, from an image of the subject's face. For example, an algorithm may analyze the relative position, size, and/or shape of the eyes, nose.A newly emerging trend, claimed to achieve improved accuracies, is 3D face recognition. This technique uses 3D sensors to capture information of a face. This information is used to identify distinctive features on the face, such as the eye sockets, nose, and chin. Various algorithms are used in the face detection like viola Jones LDP etc. Viola–Jones object detection framework is the first framework to provide competitive object detection Viola–Jones requires full view frontal upright faces the entire face must point towards the camera and should not be tilted to either side. The algorithm has four stages they are Hare Feature Selection, Creating an Integral Image, Adaboost Training Cascading Classifiers. All human faces share some similar properties. These regularities are matched using Haar Features image. The integral image evaluates rectangular features in constant time, which gives them a considerable speed advantage over more sophisticated alternative features. Local binary patterns is a visual descriptor used for classification It has been found to be a powerful feature for texture classification. It Divide the examined window into cells for each pixel, compare the pixel to each of its 8 neighbours the feature vector can be processed using the support vector machine or some other machine-learning algorithm to classify images. Such classifiers can be used for face recognition and texture analysis. A useful extension to the original operator is the called as uniform pattern.

Related work

In [1]This paper presents a syllable-based approach to unsupervised pattern discovery from speech. By first segmenting speech into syllable-like units, the system is able to limit potential word onsets and offsets to a finite number of candidate locations. These syllable tokens are then described using a set of features and clustered into a finite number of syllable classes. Finally, recurring syllable sequences or individual classes are treated as word candidates. Feasibility of the approach is investigated on spontaneous American English and Tsonga language samples with promising results.

They also present a new and simple, oscillator-based algorithm for efficient unsupervised syllabic segmentation.

In[2] The task of zero resource query-by-example keyword searches has received much attention in recent years as the speech technology needs of the developing world grow. These systems traditionally rely upon dynamic time warping (DTW) based retrieval algorithms with runtimes that are linear in the size of the search collection. As a result, their scalability substantially lags that of their supervised counterparts, which take advantage of efficient word-based indices.In this paper, they present a novel audio indexing approach called Segmental Randomized Acoustic Indexing and Logarithmic time Search (S-RAILS). S-RAILS generalizes the original frame based RAILS methodology to word-scale segments by exploiting a recently proposed acoustic segment embedding technique. By indexing word-scale segments directly, they avoid higher cost frame based processing of RAILS while taking advantage of the improved lexical discrimination of the embeddings. Using the same conversational telephone speech benchmark, they demonstrate major improvements in both speed and accuracy over the original RAILS system.

In[3] Discovering the linguistic structure of a language solely from spoken input asks for two steps: phonetic and lexical discovery. The first is concerned with identifying the categorical sub word unit inventory and relating it to the underlying acoustics, while the second aims at discovering words as repeated patterns of sub word units. The hierarchical approach presented here accounts for classification errors in the first stage by modelling the pronunciation of a word in terms of sub word units probabilistically: a hidden Markov model with discrete emission probabilities, emitting the observed unit sequences. They describe how the system can be learned in a completely unsupervised fashion from spoken input. To improve the initialization of the training of the word pronunciations, the output of a dynamic time warping based acoustic pattern discovery system is used, as it is able to discover similar temporal sequences in the input data. This improved initialization, using only weak supervision, has led to a 40% reduction in word error rate on a digit recognition task.

In [4] In this paper, they describe the “Spoken Web Search” Task, which was held as part of the 2012 MediaEval benchmark evaluation campaign. The purpose of this task was to perform audio search with audio input in four languages, with very few resources being available. Continuing in the spirit of the 2011 Spoken Web Search Task, which used speech from four Indian languages, the 2012 data was taken from the LWAZI corpus, to provide even more diversity and allow for a task that will allow both zero resource “pattern matching” approaches and “speech recognition” based approaches to participate. In this paper, they summarize the results from several independent systems, developed by nine teams, analyze their performance, and provide directions for future research.

In [5] They address the problem of estimating the latent image of a static bilayer scene (consisting of a foreground and a background at different depths) from motion blurred observations captured with a handheld camera. The camera motion is considered to be composed of in-plane rotations and translations. Since the blur at an image location depends both on camera motion and depth, de blurring becomes a difficult task. They initially propose a method to estimate

The transformation spread function (TSF) corresponding to one of the depth layers. The estimated TSF (which reveals the camera motion during exposure) is used to segment the scene into the foreground and background layers and determine the relative depth value. The de blurred image of the scene is finally estimated within a regularization framework by accounting for blur variations due to camera motion as well as depth.

In [6] In this paper they present a spoken query detection method based on posterior grams generated from Deep Boltzmann Machines (DBMs). The proposed method can be deployed in both semi-supervised and unsupervised training scenarios. The DBM-based posterior grams were evaluated on a series

Of keyword spotting tasks using the TIMIT speech corpus. In unsupervised training conditions, the DBM-approach improved upon our previous best unsupervised keyword detection performance using Gaussian mixture model-based posterior grams by over 10%. When limited amounts of labelled data were incorporated into training, the DBM-approach required less than one third of the annotated data in order to achieve a comparable performance of a system that used all of the annotated data for training.

In [7] They introduce the notion of subspace learning from image gradient orientations for appearance-based object recognition. As image data is typically noisy and noise is substantially different from Gaussian, traditional subspace learning from pixel intensities fails very often to estimate reliably the low-dimensional subspace of a given data population. They show that replacing pixel intensities with gradient orientations and the _2 norm with a cosine-based distance measure offers, to some extent, a remedy to this problem. Within this framework, which they coin IGO (Image Gradient Orientations) subspace learning, they first formulate and study the properties of Principal Component Analysis of image gradient orientations (IGO-PCA). Then they show its connection to previously proposed robust PCA techniques both theoretically and experimentally. Finally, they derive a number of other popular subspace learning techniques, namely Linear Discriminant Analysis (LDA), Locally Linear Embedding (LLE) and Laplacian Eigen maps (LE). Experimental results

Show that their algorithms outperform significantly popular methods such as Gabor features and Local Binary Patterns and achieve state-of-the-art performance for difficult problems such as illumination- and occlusion-robust face recognition. In addition to this, the proposed IGO-methods require the Eigen-decomposition of simple covariance matrices and are as computationally efficient as their corresponding _2 norm intensity-based counterparts.

In [8] They present a unified model for face detection; pose estimation, and landmark estimation in real-world, cluttered images. Their model is based on mixtures of trees with a shared pool of parts; they model every facial landmark

As a part and use global mixtures to capture topological changes due to viewpoint. They show that tree-structured models are surprisingly effective at capturing global elastic deformation, while being easy to optimize unlike dense

Graph structures. They present extensive results on standard face benchmarks, as well as a new “in the wild” annotated dataset that suggests our system advances the state-of-the art, sometimes considerably, for all three tasks. Though other model is modestly trained with hundreds of faces, it compares favourably to commercial systems trained with billions of examples.

In [9] Spoken term discovery is the task of automatically identifying words and phrases in speech data by searching for long repeated acoustic patterns. Initial solutions relied on exhaustive dynamic time warping-based searches across the entire similarity matrix, a method whose scalability is ultimately limited by the O (n2) nature of the search space. Recent strategies have attempted to improve search efficiency by using either unsupervised or mismatched-language acoustic models to reduce the complexity of the feature representation. Taking a completely different approach, this paper investigates the use of randomized Algorithms that operate directly on the raw acoustic features to produce sparse approximate similarity matrices in O (n) space and O (n log n) time. They demonstrate these techniques facilitate spoken term discovery performance capable of outperforming a model-based strategy in the zero resource setting.

In [10] This paper describes a new toolkit - SCARF - for doing speech recognition with segmental conditional random fields. It is designed to allow for the integration of numerous, possibly redundant segment level acoustic features, along with a complete language model, in a coherent speech recognition framework. SCARF performs a segmental analysis, where each segment corresponds to a word, thus allowing for the incorporation of acoustic features defined at the phoneme, multi-phone, syllable and word level.

SCARF is designed to make it especially convenient to use acoustic detection events as input, such as the detection of energy bursts, phonemes, or other events. Language modelling is done by associating each state in the SCRF with a state in an underlying n-gram language model, and SCARF supports the joint and discriminative training of language model and acoustic model parameters.

In [11] In this paper, they present an unsupervised learning framework to address the problem of detecting spoken keywords. Without any transcription information, a Gaussian Mixture Model is trained to label speech frames with a Gaussian posterior gram. Given one or more spoken examples of a keyword,

They use segmental dynamic time warping to compare the Gaussian posteriorgrams between keyword samples and test utterances. The keyword detection result is then obtained by ranking the distortion scores of all the test utterances. They examine the TIMIT corpus as a development set to tune the parameters in their system, and the MIT Lecture corpus for more substantial evaluation. The results demonstrate the viability and effectiveness of our unsupervised learning framework on the keyword spotting task.

In[12] They present a technique to automatically discover the (word-sized) phone patterns that are present in speech utterances. These patterns are learnt from a set of phone lattices generated from the utterances. Just like children acquiring language, their system does not have prior information on what the meaningful patterns are. By applying the non-negative matrix factorization algorithm to a fixed-length high-dimensional vector representation of the speech utterances, decomposition in terms of additive units is obtained. They illustrate that these units correspond to words in case of a small vocabulary task. Their result also raises questions about whether explicit segmentation and clustering are needed in an unsupervised learning context .

General

The detection of micro aneurysm has been under study for a long time. The process has to be proceeded so that there are no noises in the images and the result has to be accurate. The input image has to be processed using several techniques which were taken from various studies.

Speech to Text

In settings where only unlabelled speech data is available, speech technology needs to be developed without transcriptions, pronunciation dictionaries, or language modeling text. The trained speech is converted to text by comparing the speech with already trained data. If the speech is matched then its word will be displayed as text.

Segmentation

Speech segmentation is the process of identifying the boundaries between words, syllables, or phonemes in spoken natural languages.The term applies both to the mental processes used by humans, and to artificial processes of natural language processing.HMMs (hidden markov model) lay at the heart of virtually all modern speech recognition systems and although the basic framework has not changed significantly in the last decade or more, the detailed modeling techniques developed within this framework have evolved to a state of considerable sophistication.Various forms of GMM feature extraction are outlined, including methods to enforce temporal smoothing and a technique to incorporate a prior distribution to constrain the extracted parameters. These parameters are used to remove the noises.The feature extractor plus the SOM behaved like a transducer, transforming a sequence of speech samples into a sequence of labels.

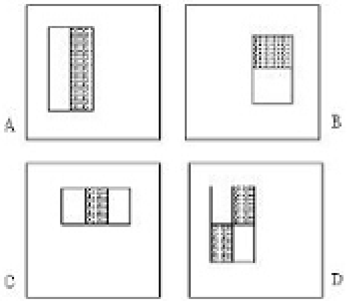
Face Detector

Viola-Jones

Viola-Jones Face Detector has three distinguished key contributions Integral Image, variant AdaBoost learning algorithm and Cascade structure to achieve high processing speed and detection rates.The detection rates of Viola-Jones face detector are comparable to the best previous systems. So it has been used in real-time applications.

Haar-like Features

They used three kinds of features: two-rectangle feature, three rectangle feature and four-rectangle feature.The value of a Haar-like feature is the difference from the sum of pixels in the grey rectangles subtracting the sum of the pixels in the white rectangles.



The grey or white regions have the same size and shape in one feature. Given a detector with the base resolution 24x24, the exhaustive set of features is over 180,000.

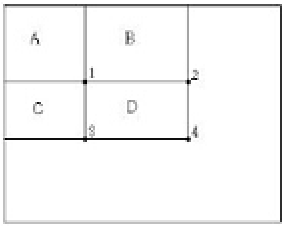
Integral Image

In order to compute these Haar-like features very rapidly at many scales the integral image representation is introduced. Any Haar-like feature can be computed at any scale or location very quickly. The integral image at location x, y contains the sum of the pixels above and to the left of x, y in the original image as bellows.

$$ii(x,y) = \sum_{x' \le x, y' \le y} i(x', y') \qquad \text{...(1)}$$

where ii is the integral image and i is the original image.

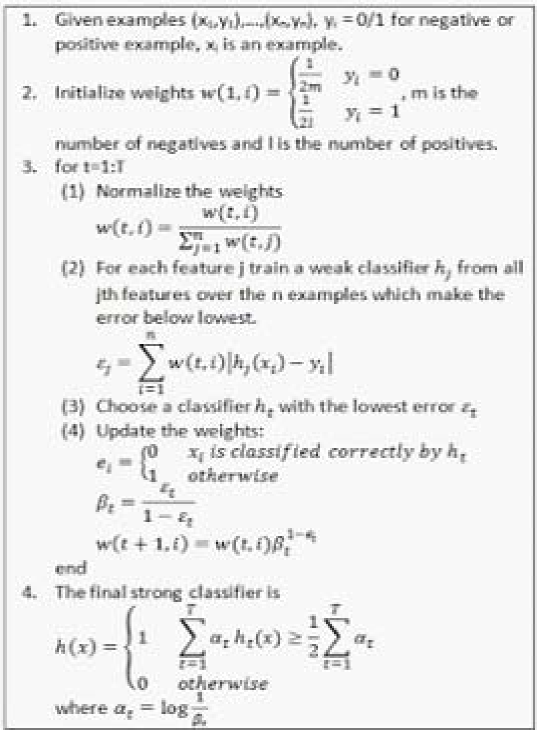
Any rectangular sum can be computed in four array references. Like the example in bellowing figure, to compute the sum of D which is equal to 4 + 1 - 2 - 3, so we need the value at 1, 2, 3 and 4 in integral image.



Learning Algorithm based on Abdboost

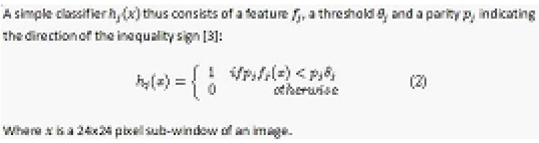
Learning Algorithm based on Abdboost

Within any image subwindow the total number of Haar-like features is very large. In order to ensure fast classification, the learning process must focus on a small set of critical features. Here, a modified AdaBoost is used not only to select a small set of features but also to train the classifier as below.



```
1. Given examples ($x_1$,$y_1$),...,($x_n$,$y_n$), $y_i$ = 0/1 for negative or
   positive example, $x_i$ is an example.
2. Initialize weights $w(1,i) = \begin{cases} \frac{1}{2m} & y_i = 0 \\ \frac{1}{2l} & y_i = 1 \end{cases}$, m is the
   number of negatives and l is the number of positives.
3. for t=1:T
   (1) Normalize the weights
       $w(t,i) = \frac{w(t,i)}{\sum_{j=1}^{n} w(t,j)}$
   (2) For each feature j train a weak classifier $h_j$ from all
       jth features over the n examples which make the
       error below lowest.
       $\varepsilon_j = \sum_{i=1}^{n} w(t,i)|h_j(x_i) - y_i|$
   (3) Choose a classifier $h_t$ with the lowest error $\varepsilon_t$
   (4) Update the weights:
       $e_i = \begin{cases} 0 & x_i \text{ is classified correctly by } h_t \\ 1 & otherwise \end{cases}$
       $\beta_t = \frac{\varepsilon_t}{1 - \varepsilon_t}$
       $w(t+1,i) = w(t,i)\beta_t^{1-e_i}$
   end
4. The final strong classifier is
   $h(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \ge \frac{1}{2}\sum_{t=1}^{T} \alpha_t \\ 0 & otherwise \end{cases}$
   where $\alpha_t = \log\frac{1}{\beta_t}$
```

Using the AdaBoost we need a simple learning algorithm which selects the single rectangle feature which best separates the positive and negative examples.



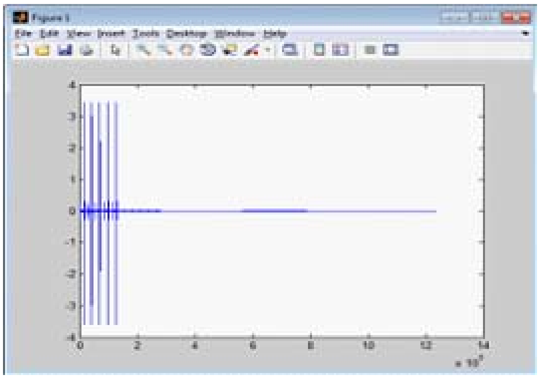
A simple classifier $h_j(x)$ thus consists of a feature f_j, a threshold θ_j and a parity p_j indicating the direction of the inequality sign [3]:

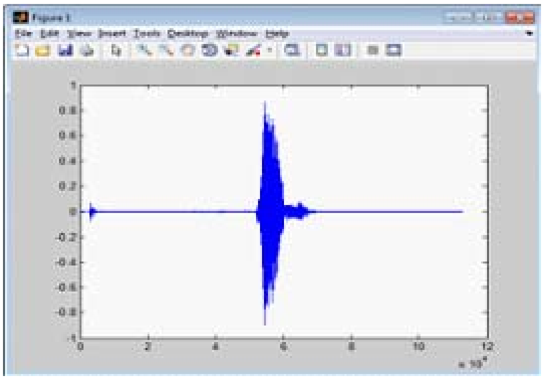
$$h_j(x) = \begin{cases} 1 & if\, p_j f_j(x) < p_j \theta_j \\ 0 & otherwise \end{cases} \qquad (2)$$

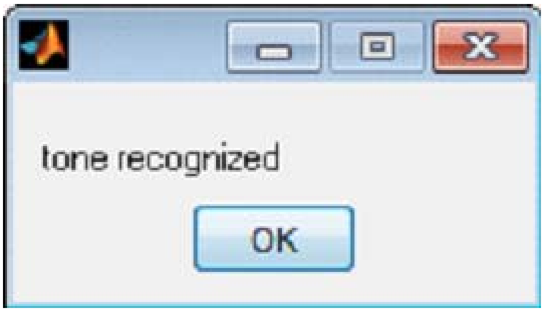
Where x is a 24x24 pixel sub-window of an image.

Cascade Structure

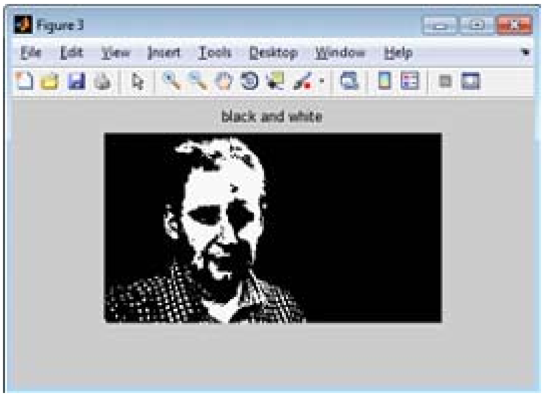
Cascade structure increases the speed of the detector by focusing on face-like regions of the

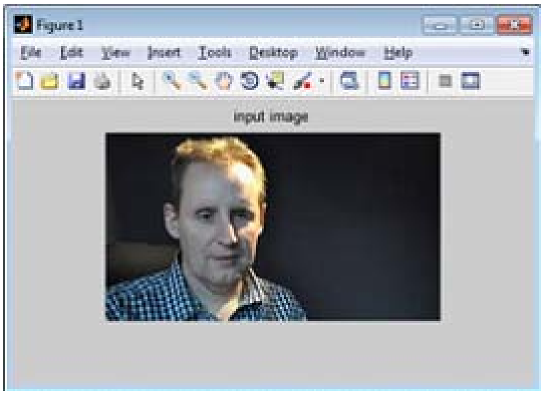






tone recognized





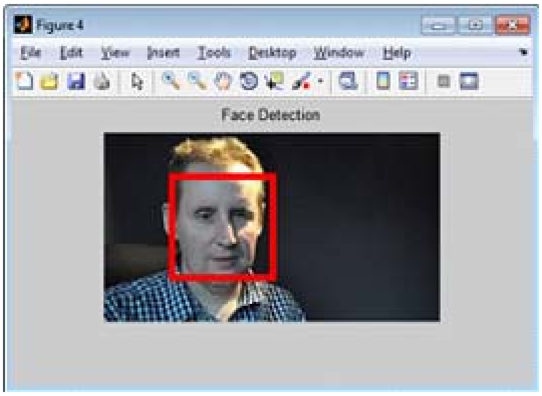


image.It is a sequence of more and more complex classifiers that sub-windows which are not rejected by the last classifier will be processed by the next more complex classifier.

The cascade structure contains many classifier stages. In each stage there is a strong classifier trained by using modified AdaBoost on increasingly features until the target detection and false positives rates are met.In the test of original Viola-Jones Detector, the cascade has 38 stages with over 6000 features.

CONCLUSION

An efficient application has been proposed to perform voting by blind people. A HMM (Hidden Markov Model), GMM (Gaussian Markov Model) and SOM (self organized mapping) segmentation are used to analyze the system in CFCC and MFCC with SVM trained samples of speech. Viola Jones algorithm is used for face detection and SVM classifier is used to recognize the face. Efficient outcome is obtained when compared to existing HMM model for speech.

Future enhancements

In proposed work a simple idea has been proposed for developing a system for blind people to help them in voting easily. In future we will develop a system with more efficient camera and better techniques for speech recognition.

REFERENCES

1. Okko Rasanen, Gabriel Doyle, Michael C. Frank Unsupervised Word discovery from speech using automatic Segmentation into syllable-like units “, *Int. Jour. Computer Applications*, vol. 117, no. 1, May 2015.
2. Keith Levin, Aren Jansen, Benjamin Van Durme” Segmental acoustic indexing for zero resource keyword search” Human Language Technology Center of Excellence, Center for Language and Speech Processing Johns Hopkins University, Baltimore 2015
3. Oliver Walter, Timo Korthals and Reinhold Haeb-Umbach A Hierarchical system for word discovery exploiting dtw-based initialization”2015.
4. Vijay Rengarajan1 , Abhijith Punnappurath1 , A.N. Rajagopalan1 , and Guna Seetharaman2 “Efficient Change Detection for Very Large Motion Blurred Images “Efficient Change Detection for Very Large Motion Blurred Images
5. Stephen J. Wright, Dimitri Kanevsky, Li Deng, Xiaodong He, Georg Heigold , and Haizhou Li,,” optimization Algorithms and Applications for Speech and Language Processing,” *IEEE Transactions on audio, speech, and language processing*, **21**(11): (2013)
6. Todd K. Moon, Jacob H. Gunther, Cortnie Broadus, Wendy Hou and Nils Nelson,” Turbo Processing for Speech Recognition,*” IEEE transactions on cybernatics,* **44**(1) (2014).
7. Tianyu , T. Wang, and Thomas F. Quatier,” Towards Interpretive Models for 2-D Processing of Speech,”*IEEE transactions on speech language and processing,* **20**(7) (2012).
8. Simon Lucey, Tsuhan Chen, Sridha Sridharan and Vinod Chandran,” Integration Strategies for Audio-Visual Speech Processing: Applied to Text-Dependent Speaker Recognition,” IEEE *Transactions On Multimedia*, **7**(3); (2005).
9. Mari Ostendorf, Benoit Favre, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Dustin Hillard, Julia Hirschberg, Heng Ji, Jeremy G. Kahn, Yang Liu, Sameer Maskey, Evgeny Matusov, Hermann Ney, Andrew Rosenberg, Elizabeth Shriberg, Wen Wang, and Chuck Wooters,” Speech Segmentation and Spoken Document Processing,” *IEEE Signal processing magazine* (2008).
10. Lan, K. B. Nie, S. K. Gao, and F. G. Zeng,” A Novel Speech-Processing Strategy Incorporating Tonal Information for Cochlear Implants,” *IEEE Transactions on biomedical engineerinG*, **51**(5): (2004).
11. Alex S.Park, James R.Glazz “Unsupervised Pattern Discovery in Speech,” *IEEE Transactions on audio, speech, and language processing,* **16**(1), JANUARY 2008.
12. Dimitrios Dimitriadis and Enrico Bocchier,” Use of Micro-Modulation Features in Large Vocabulary Continuous Speech Recognition Tasks,” IEEE/ACM Transactions on audio, speech, and language processing, **23**(8), august 2015.