International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 4, June - 2012

Mathematical Model Based on Human Speech Recognition and Body Recognition

[1]Sneha K. Patel
Assistant Professor,
Humanities and Social Science department
SSASIT, Surat

[2]Dr. J. M. Dhodiya
Assistant Professor,
Department Of Applied Mathematics & Humanities
SVNIT, Surat.

[3]Dr. D. C. Joshi
Associate Professor
Department of Mathematics
VNSGU, Surat.

Abstract

Technology is the fastest growing field now-a-days. In this vast field one must need security system which uses electronics gadgets. And this security system must be growing and updating simultaneously with technology. This growth in electronic transactions results in a raise of demand for fast and accurate user identification and authentication system. Total security system may solve this problem since number of parameters like face, speech, fingerprint, palm print etc. are undeniably connected to its owner. It is also verify quantitative data like E-cards, password and Login ID etc. of human being. In this paper we have discussed the security system based on speech recognition and body recognition of human kind through mathematical model. This proposed model for the system can compare the recorded speech and body expressions with original speech and body expressions which is stored in a central or local database to give perfect identification.

Key words: Security, Qualitative and Quantitative data, GMM, Speech recognition, Body recognition.

1. Introduction

Technology is growing up day by day in the present era. The rapid growth in the use of internet applications and the great concern for security require reliable and automatic personal identification. In this vast field one must need security system which uses electronics gadgets/devices and this security system must be grown up and updated simultaneously with technology. This growth in electronic transactions results in a raise of demand for fast and accurate user identification and authentication system. Total security system may solve this problem since number of parameters like face, speech, fingerprint, palm print etc. are undeniably connected to its owner and it is also verify quantitative data like E-cards, password and Login ID etc. of human being.

The aim of this paper work is to explore speech recognition and body recognition using mathematical techniques and its applications for security system. In our minds the aim of interaction between a machine and a human is to use the most natural way of expressing ourselves, through our speech and body. Speech recognition, which can be classified into identification and verification, is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. Speech verification is the process of accepting or rejecting the identity claim of a speaker. Most applications in which a speech is used as the key to confirm the identity of a speaker are classified as speaker verification [25]. But sometimes speech has the problem of not being able to identify human with sore throat. In this type of problem we solve with body recognition.

In body recognition, digital video cameras connected to computers have come into wide use recently. Body recognition has to be able to perform the basic two tasks: (a) Detect and track people and (b) Person recognition. Today there are a number of computer techniques that can be used in automatic visual surveillance systems: Face detection [3, 9, 10, 12] and face recognition [19, 1, 14] have been thoroughly studied over the last 10 years in computer field. While recent face detection systems [3, 10] are able to deal with large pose variations, face recognition systems are limited to identifying persons in frontal and near-frontal views only. Recently, learning-based techniques [5] and template matching [6] have been applied to detecting people in still images. As shown in Refs. [20, 2] the periodicity of gait allows to detect walking people in image sequences. Gait has also been used for person recognition in image sequences [15]. There is number of challenges for human body detection and identification like the invariance against pose changes, changes in illumination, and the selection of image features which allow to reliable identify human body.

In this paper researchers discussed speech recognition and body recognition for security purpose with Gaussian Mixture Models (GMM) and Body recognition Model.

2. Objectives

- To use Speech as the key to confirm the identity and verification of a speaker. As Speech recognition technique makes it possible to use the speaker's speech to verify their identity and control access to services such as voice dialing, banking by telephone, database access services, information services, voice mail, telephone shopping, security control for confidential information areas, and remote access to computers [24].
- To use Body expression recognition for verification of human for total security system with mathematical algorithms.

Speech recognition is generally used as a human – E-machine interface for other software. A speech recognition system performs three primary tasks:

- Preprocessing – Converts the spoken input into a form the recognizer can process.
- Recognition – Identifies what has been said.
- Communication – Sends the recognized input to the software/hardware systems that need it.

Speech recognition is the translation of spoken words into text. There are two types of speech recognition one text-dependent and other text-independent and different methods likes HMMs, GMMs, SVMs and NNs may be used for it. The HMM (Hidden Markov Models) is usually for text-dependent speaker recognition since there is textual context. As a special case of HMM, GMM are used for doing text-independent voice recognition. Here we see GMM (Gaussian Mixture Models) for speech recognition.

3. Gaussian Mixture Models

GMM is a popular technique to represent the speaker, which is a commonly used estimate of the probability density function. The Gaussian mixture speaker model was introduced in [21], and has demonstrated high text- independent recognition accuracy for short test sounds. To build a Gaussian Mixture Model of a speaker's voice, one should make a few assumptions and decisions. The first assumption is the number of Gaussians to use. Gaussians is a set of parameters, each specifying the parameter of the corresponding mixture component. This is dependent on the amount of data that is available and the dimensionality of the feature vectors. Once the number of Gaussians is determined, some large group of features is used to train these Gaussians. This step is said to be training. The models generated by training are called universal background models [11].

In GMM a non-singular multivariate normal distribution of a d- dimensional random variable x can be defined as

$$\left.\begin{aligned} P(x) &= \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma|}} e^{\left\{-\frac{1}{2}(x-\mu)^T(\Sigma^{-1})(x-\mu)\right\}} \\ \text{where } &\begin{cases} x,\ \mu \in \mathcal{R}^d \\ \boldsymbol{\Sigma}:\ \mathcal{R}^d \to \mathcal{R}^d \end{cases} \end{aligned}\right\} \quad (1)$$

When observations are made on more than one variable then it is called Multivariate data. In Equation 1, $P(x)$ is called Probability Density Function (PDF) formula, μ is the mean vector $(d \times 1 \text{ matrix})$ and $\boldsymbol{\Sigma}$ is the covariance matrix $(d \times d \text{ matrix})$ of the normally distributed random variable x. In Equation 1, we are take $d = 2$-dimensional in Gaussian Probability Density Function (PDF) formula.

In Equation 1, μ is the mean vector (Expected value) which defined as

$$\mu \triangleq F(x) \triangleq \int_{-\infty}^{\infty} x\, P(x)\, dx$$

(2)

The so-called "Sample Mean" approximation for Equation 2 is

$$\mu \approx \frac{1}{N}\sum_{i=1}^{N-1} x_i \quad (3)$$

Where N is the number of samples and x_i are the Mel-Cepstral feature vectors [8]. The main aim of Mel-Cepstral feature vector is to capture important information presented in a speech signal for recognition purpose and also use for fast evaluation algorithm for speech recognition. The variance-Covariance matrix of a multi-dimensional random variable is defined as,

$$\Sigma \triangleq F\{(x - F\{x\})(x - F\{x\})^T\} \quad (4)$$

$$= F\{xx^T\} - \mu\mu^T \quad (5)$$

This matrix is called the Variance-Covariance since the diagonal elements are the variances of the individual dimensions of the multi-dimensional vector x. The off-diagonal elements are the Covariances across the different dimensions. Here we said to be this matrix is Covariance matrix. The unbiased estimate of $\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}$ is given by the following expression,

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{i=1}^{N-1}(x_i - \mu)(x_i - \mu)^T \quad (6)$$

$$= \frac{1}{N-1}\left[S_{xx} - N(\mu\mu^T)\right] \quad (7)$$

Where the sample mean μ is given by Equation 3 and the second order sum matrix S_{xx} is given by

$$S_{xx} = \sum_{i=1}^{N-1} x_i x_i^T \quad (8)$$

Now the training data is ready and the basis for a speech independent model is built which is stored in the form of the above statistics. For a UBM, a set of speakers is used to optimize the parameters of the Gaussians as well as the mixture coefficients, using standard techniques such as maximum likelihood estimation (MLE), Maximum a Posteriori (MAP) adaptation and Maximum Likelihood Linear Regression (MLLR). At this point, the system is ready for performing the enrollment. The enrollment may be done by taking a sample audio of the target

voice and adapting it to be optimal for fitting this sample. This ensures that the likelihoods returned by matching the same sample with the modified model would be maximal.

4. Body Recognition

Images of full-body persons are represented by color-based and shape-based features. The system consists of two modules one is Image pre-processing and other is human body/pose recognition. Overview of the human body recognition is shows in Figure 1.

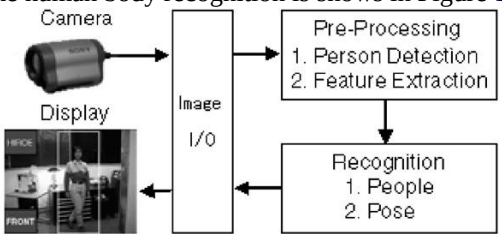

Figure 1. The system overview of human body recognition.

First of all, human image is capture from the camera and forwarded to the pre-processing module, where human body and its feature, extract from the background. In recognition module, human body is recognized with the human identity and pose of human.

4.1 Pre-processing

The pre-processing module consists of two parts: One human body detection of moving persons and other extraction of image of human body features [4].

Human body detection:

The system uses two steps to detect a moving body of human in an image sequence. In the first step the system subtracts the current background image from the latest k images, and stores one of these k images. Here we take $k = 3$. But if human body image has energy larger than a threshold then the result of background subtraction may include a lot of noise and therefore the stored image does not contain a human body. In this case, the second step helps to remove those images, which not containing a human body. For this purpose, the system extracts the shape of a possible human body by using edge detection. We assuming that the human body is slightly moving between two frames, the system performs edge detection on the image obtained by subtracting two consecutive images in the sequence. If the number of edge pixels is larger than a threshold, one of the k images is eventually stored. Finally, if no person image is detected, the background is updated by computing the average of the k latest images. We can see in Figure 2(a) shows an image from the sequence and in figure. 2(b) shows the combined result of the two steps.

Figure 2. Moving human body detection.

Feature extraction:

In image processing and photography, a color histogram is a representation of the distribution of colors in an image. For digital images, a color histogram represents the number of pixels that have colors in each of a fixed list of color ranges that span the image's color space, the set of all possible colors. All images must be represented in matrix form. Once the human body has been detected and extracted from the background, we calculate different types of human body image features:

(1) RGB color histogram:

We calculate one dimensional color histogram with 32 bins for each color channel. Where bins means all elements in vector Y or in one column of matrix Y are grouped according to their numeric range. Each group is shown as one bin. Here the total number of extracted features is 96 (32 × 3) for a single image.

(2) Normalized color histograms:

We calculate two dimensional normalized color histograms; r=R/(R+G+B), g=G/(R+G+B). Again, we chose 32 bins for each color channel. Overall, the System extracts 1024 (32 × 32) features from a single image.

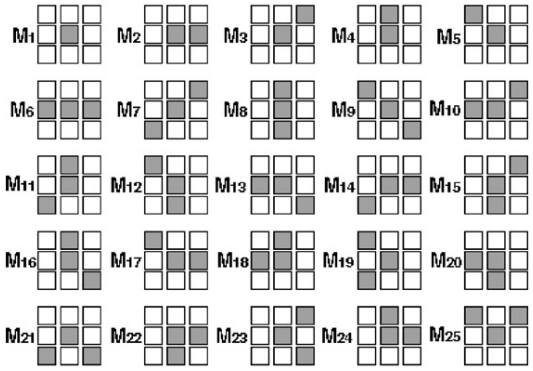

Figure 3. Shape patterns

(3) RGB color histogram + shape histogram:

Image is stored in rows and columns form. We calculate simple shape features of people by counting pixels along rows and columns of the extracted human body images. We choose a resolution of 10 bins for column histograms and 30 bins for row histograms. The total number of extracted features is 136, 32×3 for the RGB histograms and 10 + 30 for the shape histograms.

(4) Local shape features:

Local features of an image are obtained by convolving the local shape patterns shown in Figure 3. These patterns were introduced in Ref. [23] for position invariant human body detection. Let $M_i ; i = 1,2, \dots 25$, be the patterns in Figure 3 and V_k the 3×3 patch at pixel k in an image. We consider two different types of convolution operations. The first is the linear convolution given by $\sum_k V_k \cdot M^i$, where the sum is on the image pixels. This pattern is shows in Figure 3 from 1 to 5. The second is a non-linear convolution given by

$$F_i = \sum_k C_{(ki)} \quad \text{Where}$$

$$C_{(ki)} = \begin{cases} V_k \cdot M^i & \text{if } V_k \cdot M^i = \max_j (V_k \cdot M^j) \\ 0 & \text{otherwise.} \end{cases}$$

It is shows in Figure 3 from 6 to 25. The non-linear convolution mainly extracts edges and has been inspired by recent work in the field of brain models [23]. The shape features are extracted for each of the following color channels separately: $R + G - B$, $R - G$ and $R + G$. This color model has been suggested by physiological studies [13]. The system extracts 75 (25 × 3) features from the three color channels.

4.2 Recognition

We first collect an N set of data images of human body and manually gives label to it, according to the identity and pose of the human body. The pose (right, left, front and back) of the human body are stored with label in database. The set of input–output is denoted as $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, where the input x_i denotes the feature vector extracted from image i and the output y_i is a class label. We use Support Vector Machine (SVM) classifiers for human body recognition. The standard SVM takes a set of input data and predicts for each given input. Here we train different SVM classifiers on the labeled data to perform the multi-class classification task for human body identification and pose estimation. SVMs are a technique to train classifiers and probability densities that is well-founded in statistical learning theory [26]. One of the main attractions of using SVMs is that they are capable of learning in sparse, high dimensional spaces with very few training examples. SVMs accomplish this by minimizing a bound on the experimental error and the complexity of the classifier, at the same time. This controlling of both the training error and the classifier's complexity has allowed SVMs to be successfully applied to very high dimensional learning tasks such as face detection [7], 3-D object recognition [16], stop word detection in speech signals [18], and text categorization [22]. We will apply SVMs to very high dimensional classification problems.

5. Applications

The present era of information and technology is quickly revolutionizing the way of transactions. And security plays an essential role in technology. For example access codes for banks accounts and computer systems often use PIN's for identification and security clearances. When credit and ATM cards are lost or stolen, an unauthorized user can often come up with the correct personal codes. In this case Total Security System may solve this problem since the Total Security System works with qualitative data and quantitative data of human and identifies human characteristics, like face, speech, fingerprint, palm print, body etc. are undeniably connected to its owner. This system is highly beneficial for Bank, Military, Crime branch etc.

As described above GMM and SVM used for voice recognition and human body recognition respectively and so both recognition techniques are most useful in total security system. The total security system is the system, which is verifies quantitative data like E-cards, login ID and password as well as qualitative data like face, voice, body, Iris of human.

The advantages of using a GMM as the likelihood function are that it is computationally inexpensive, is based on a well-understood statistical model. It is insensitive for text-independent tasks of the voice from the database.

Let $X_1, X_2, \dots, X_n$ have Probability Density Function (PDF)

$$P(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_m)$$

where the parameters $\theta_1, \theta_2, \dots, \theta_m$ have unknown values. When $x_1, x_2, \dots, x_n$ are the observed sample values and P is regarded as a function of $\theta_1, \theta_2, \dots, \theta_m$, which is called likelihood function. The maximum the likelihood estimates $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m$ are those values of the θ_is that maximize the likelihood function, so that

$$P(x_1, x_2, \dots, x_n; \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m) \geq P(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_m) \quad ; \forall \theta_1, \theta_2, \dots, \theta_m$$

.

when the X_is are replace with x_is, which gives maximum likelihood estimation output. Thus, electronic device identify highest likelihood value from the collection of training data, which data is generated by the GMM. To ensure a good dynamic range and better discrimination capability, log of the likelihood is computed. At the verification stage, the process is very similar to the identification process described above, with the exception that instead of computing the log likelihood for all the voices in the database. Input voice is compared with the database. If the input voice gives a better log likelihood, the input voice (speaker) is verified and otherwise rejected. The comparison is done using the Log Likelihood Ratio (LLR) test. When the logarithm of

the likelihood ratio is used, the statistic is known as a log-likelihood ratio statistic.

In body recognition, the system recorded human expression and pose during whole one day. From the video recording, system capture different poses(left, right, front, back) and expression of the one person and prepare the training data with normalizing color, features and expressions. Now whenever next time person is passing through the system, the pose and expressions are captured. The systems first recognize the person and then select the proper multi-class classifier (SVMs), to determine the pose of the person. The system returns the pose identification rates and human expression identification rates of the input person. If it is match from the training data then person is verified otherwise reject. For both tasks, person identifications and pose estimation the best results are obtain with normalized color features as well as SVM classifiers.

6. Conclusion

The Total security system is a proposed system work with speech recognition and body expression. We have presented a model that recognizes speech and body of a human from environment. We have used GMM for speech recognition and for human body recognition performs by multi-class SVMs. The proposed system works in real time using SVMs, achieving high recognition rate on normalized color features and poses of body. This work can be readily used in biometric applications like access control and verification systems.

7. References

[1] A.Pentland, T.Choudhury, *Face recognition for smart environments*, *Computer* 33(2)(2000)50–55.
[2] B.Heisele, C. Woehler, "Motion-based recognition of pedestrians", *Proceedings of International Conference on Pattern Recognition and Image Processing*, 1998, pp. 1325–1330.
[3] B.Heisele, T. Poggio, M. Pontil, *Face detection in still gray images*, MIT AI Memo 1687, 2000.
[4] C.Nakajima, M. Pontil, B. Heisele,T. Poggio, *The Journal of the Pattern recognition socirty*,2003
[5] C.Papageorgiou, T. Poggio, *A trainable object detection system: car detection in static images*, MIT AI Memo 1673, 1999.
[6] D.avrila, *Pedestrian detection from a moving vehicle, Computer Vision: 3rd European Conference on Computer Vision*, 2000, pp. 37–49.
[7] E.Osuna, R. Freund, F. Girosi, "An improved training algorithm for support vector machines", *Proceedings of IEEE Workshop on Neural Networks and Signal Processing*, 1997, pp. 276–285.
[8] H.Beigi, *Fundamentals of speaker Recognition, Springer*, New York, ISBN:970-0-387-77591-3,2011.
[9] H.Rowley, S. Baluja, T. Kanade, "Neural network-based face detection", *IEEE Trans. Pattern Anal. Machine Intelligence* 20 (1) ,1998 pp. 23–38.
[10] H.Schneiderman, T. Kanade, "A statistical method for 3d object detection applied to faces and cars", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*,2000, pp.746–751.
[11] J.Yang. *Biometrics*, InTech, ISBN 978-953-307-618-8
[12] K.Sung, T. Poggio, *Example-based learning for view-based human face detection*, MIT AI Memo 1521, 1994.
[13] K.Uchikawa, *Mechanism of color perception*, Asakura syoten, 1998.
[14] L.Wiskott, J.M. Fellous, N. Kruger, C. von der Malsburg "Face recognition by elastic bunch graph matching", *IEEE Trans. Pattern Anal. Machine Intelligence* 19 (7),1997, pp. 775–779.
[15] M.Nixon, J.N. Carter, J.M. Nash, P.S. Huang, D. Cunado, S.V. Stevenag, "Automatic gait recognition", *IEE Colloq. Motion Anal. Track. 3*, 1999, pp. 1–6.
[16] M.Pontil, A. Verri, "Support vector machines for 3-d object recognition", *IEEE Transactions on Pattern Analysis Machine Intelligence* 20 (6), 1998, pp. 637–646.
[17] M.Riesenhuber, T. Poggio, *Hierarchical models of object recognition in cortex*, Nature Neurosci. 2, 1999, pp. 1019–1025.
[18] P.Niyogi, C. Burges, P. Ramesh, "Distinctive feature detection using support vector machines", *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, 1999, pp. 425–428.
[19] R.Brunelli, T. Poggio, "Face recognition: Features versus templates", *IEEE Trans. Pattern Anal. Machine Intelligence* 15 (10), 1993, pp. 1042–1052.
[20] R.Cutler, L. Davis, "Robust periodic motion and motion symmetry detection", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2000, pp. 615–622.
[21] R.Duda, P. Hart and D. Stork, *Pattern Classification*, 2nd ed. ,2001,Wiley, NewYork.
[22] T.Joachims, Text categorization with support vector machines, Technical Report LS-8 Report 23, University of Dortmund, 1997.
[23] T.Kurita, K. Hotta, T. Mishima, "Scale and rotation invariant recognition m ethod using higher-order local autocorrelation features of log-polar image", *proceedings of IEEE Asian Conference on Computer Vision*, Vol. 2, 1998, pp. 89–96.
[24] T.Matsui, and S. Furui, "Concatenated phoneme models for text-variable speaker recognition", *Proceedings of ICASSP'93*, 1993, pp. 391-394.
[25] T.Matsui, and S. Furui, "Similarity normalization method for speaker verification based on a posteriori probability", *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, 1994, pp. 59-62
[26] V.Vapnik, *Statistical Learning Theory*, Wiley & sons, Inc., New York, 1998.