Turkish Journal of Computer and Mathematics Education *Vol.12 No.3(2021), 3395-3404*

Research Article

Performance Evolution of Face and Speech Recognition system using DTCWT and MFCC Features

Shanthakumar H.C[a], Nagaraja G.S[b], Mustafa Basthikodi[c]

[a] Department of Computer Science Engineering, SJBIT, (Research Scholar, Jain University) Bengaluru, India
[b] Department of Computer Science Engineering, RV College of Engineering (IEEE Senior Member), Bengaluru, India
[c] Department of Computer Science Engineering, Sahyadri College of Engineering & Management, Mangaluru, India

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: Every activity in day-to-day life is required the need of mechanized automation for ensuring the security. The biometrics security system provides the automatic recognition of human by overcoming the traditional recognition methods like Password, Personal Identification Number and ID cards etc. The face recognition is a wide research with many applications. In the proposed work face recognition is carried out using DTCWT (Dual Tree Complex Wavelet Transform) integrated with predominant QFT (Quick Fourier Transform) and speech recognition is carried out using MFCC (Mel Frequency Cepstral Coefficients) algorithm. The distance formula is used for matching the test features and database features of the face and speech images. Performance variables such as EER, FRR, FAR and TSR are evaluated for person recognition

Keywords: DT-CWT, QFT MFCC, Euclidean Distance, Face Recognition

1. Introduction

In the recent years, with the drastic development in computer technology, the security methods for authentication have switched from traditional methods like Identity Card, PIN and Password etc to biometrics, as the biometric security methods are convenient for users and tough to steal and breach.

Biometric is a Greek term in which Bio stands for life & metric stands for measure that means measurement of life [1]. Biometric works on two characteristics such as physiological and behavioral. The physiological characteristics are based upon the shape of human body like face, iris, palm print and fingerprint whereas the behavioral characteristics of a person based on his behavior, which include signature, gait and voice, where face detection is one among the most commonly used biometric for verification in security systems as it is intuitive and non-intrusive. It is one of dynamic area of research, with a specialized application in various fields. It has wide-ranging application prospects in many fields like automation access control, video content indexing, surveillance, crime investigation, human computer interaction and other fields.

The main form of communication between human beings is speech & recognition of speech made possible for machine to understand human language. Automatic Speech Recognition (ASR) technology has decreased the human efforts in different fields by using an efficient user interface for various devices in all applications of computer technology like telephone networks, voice dictation, voice navigation in smart phones, smart speakers etc. Speech recognition system recognizes the input as language of human through device and analyzes the language of the human and then transforms the voice signal in the process into the corresponding logical information that can be recognized by the computer [1]. In speech recognition method pre-processing is carried out to obtain the speech signal and features are extracted to match the recognized speech signal with test signal using distance formula.

2. Related Works

Tadi Chandrasekhar and Ch. Sumanthkumar [2] proposed a model for recognizing the face by using Adaptive Neuro- Fuzzy Interface System classifier. In this model DTCWT was utilized for enhancing face pictures. The discriminative face features of these enhanced images are extracted using Principal Component Analysis (PCA) method. Performance of suggested approach were measured on YALE B as well as ORL data sets with ANFIS classifier. Researchers in [3] designed an algorithm using fast PCA & HOG (Histogram of Oriented Gradient) for recognizing the face under non-restrictive environment. Preprocessing of the raw data was carried out to extract the face region using Haar feature classifier. HOG features of this face image are extracted. The experimentation was conducted using Support Vector Machine (SVM) for matching on Labeled Faces in the Wild (LFW) database. Ningning Zhou et al., [4] constructed a face recognition system technique by improving the CS-LBP (Center-

Symmetric Local Binary Pattern). In CSLBP ignoring of the central pixel information has an impact on discriminative ability. To overcome this effect a descriptor was designed for feature extraction by fusing the central pixel information into CS-LBP. The effectiveness of the algorithm was tested on data sets like FERET, YALE B, YALE, and ORL by adopting the nearest neighbor classifier for matching process. An algorithm was developed by Ravi J et al., [5] on the basis of DTCWT and LBP for face recognition. The original face images of all the databases are resized for uniformity in the preprocessing stage. The DTCWT coefficients of the resized face images are extracted using five levels DTCWT, which are then segmented into 3X3 matrix. The final face features are extracted from segmented matrix using the LBP descriptor. The experiment was carried out on different databases by comparing the test features with the trained features using Euclidean Distance classifier.

In [6] the researcher developed a face recognition algorithm using fusion of feature learning techniques. In this technique desired face region was captured by tree structure part model on the basis of facial landmark points. From these face region patches Scale Invariant Feature Transform (SIFT) descriptors were determined. Feature learning method such as block co-ordinate decent, sparse representation coding, co-ordinate decent, locality constraint linear coding is applied on SIFT descriptors for obtaining different input image face features. Such scores of this learning technique are fused to make a decision in recognition process. The performance is evaluated on different databases using Multiclass SVM. Lijian Zhou et al., [7] have designed an algorithm on the basis of 2DLPP (two-dimensional locality preserving projection) & LBP for face recognition. Enhanced texture features are extracted using LBP descriptor by eliminating the illumination and noise effects, then 2DLPP was applied on these enhanced features to capture some features and by reducing dimension of image space structure data. In order to assess the effectiveness of algorithm, Experiment were performed on Yale, the expanded CMU PIE C09 & Yale B standard database with Nearest Neighborhood Classifier (NNC). Tong Xiaoet al., [8] has developed the technique of encrypted face recognition using Tent Map, Discrete Cosine Transform, Discrete Wavelet Transform. A pseudo-random sequence was generated with the use of Tent Map. DWT-DCT was applied on the face image to extract coefficient matrix, dot product of these matrix was done with pseudo-random sequence to get the encrypted face image. Projection matrix was generated from the encrypted image by the application of PCA, which was used to train the Back Propagation neural network for recognition. The simulation experiment was conducted on ORL database to check the robustness of the algorithm.

Eyad I. Abbasand Mohammed E. Safi [9] developed the algorithm for face recognition by reducing the database size by wavelet decomposition. Discrete wavelet decomposition was applied on the training and test images to decrease the database size. Final features are extracted from these images using PCA. The algorithm was tested on ORL database using Euclidean Distance classifier.

Punnam and Satyasavithri [10] proposed DT-CWT sub band segmenting for recognizing face. With the use of DT-CWT the image of the face is split into various oriented sub bands. Novel one's representation of this sub-band was formed by arranging from low to high frequency as a column vector, PCA was applied on this representation to get the final features. The Performance was evaluated on the ORL database with the use of k nearest neighbor classifier with Maholanobiscosine distance.

Hua Wanget al., [11] developed the algorithm by fusing HOG and Local Difference Binary for face recognition. Local pattern characteristics of face image is obtained through LDB descriptor and edge features are extracted by HOG descriptor. These features are fused to extract the final feature vector. The accuracy of the algorithm was tested on ORL and Yale datasets using linear SVM classifier. Chunling Tang an Min Li [12] proposed an algorithm for speech recognition in the noise environment using speech enhancement, combined with discard feature model by eliminating noise to check the correct voice in the voice information and finally measured the speech recognition rate in the noise environment of automobile. Lucas Debatinet al., [13] has proposed offline speech recognition techniques by referring different speech recognition topics. Author concluded to improve the speech recognition rate by reducing error rate, neural networks for language models and n-gram statistical models.

R. Thiruvengatanadhan [14] developed an algorithm for speech recognition using Auto associative Neural Network technique. In the proposed algorithm speech features are extracted using Mel Frequency Cepstral Coefficients (MFCC) for individual word which is trained to the system and recognition rate is measured.

Ritesh A. Magre, and Ajit S. Ghodke [15] developed robust feature extraction for visual speech and speaker recognition algorithm is proposed. In the proposed work features are extracted by considering the speakers moth region and also compared with different visual features to select the best feature to increase the accuracy and reliability for identification visual speech.

Mehryar Mohri et al., [16] proposed frame work with weighted finite-state transducers for speech recognition. Proposed speech recognition system is developed by considering different components of transducers to provide context-dependency models, statistical grammars, hidden Markov models, pronunciation dictionaries, and phone/word lattices.

Ashok Kumar and Vikas Mittal [17] developed algorithm for speech recognition using different methods like Linear predictive Cepstral coefficients, Mel Frequency Cepstral Coefficient, and PLP feature extraction methods.

Jayanthi Kumari and Jayanna [18] presented the research work with restricted data (not more than 15 seconds) using different feature extraction algorithm that provide good performance of speaker verification. In proposed work features are extracted by considering different methods such as LPCC, MFCC, LPRP and Linear Prediction for NIST-2003 database to measured equal error rate.

P. Krishnamoorthy et al., [19] has proposed a method for recognition of speaker under the condition of limited data with additional noise for obtaining better performance for limited data (less than 15 s) and measured Signal to Noise Ratio (SNR) for 100 speakers by selecting randomly in TIMIT database. To measure the performance of the system different feature extraction techniques like MFCC and Gaussian Mixture Model are used.

3. Proposed Model

The proposed work consists of necessary Dual-Tree Complex Wavelet Transform along with Quick Fourier Transform (QFT) is used to obtain the combined face features &Mel-Frequency Cepstral Coefficients are utilized for obtaining speech characteristics to recognise the person more accurately. Figure 1 shows the proposed model for face recognition.

DTCWT seems to be the competent algorithm for applying a wavelet transition. The technique is known for having Fourier Transformation properties in wavelet transform. Dual-Tree Complex Wavelet Transform algorithm has several advantages like limited redundancy & perfect reconstruction, approximate shift invariance directional selectivity in addition to that greater basis for de-blurring & de-noising.

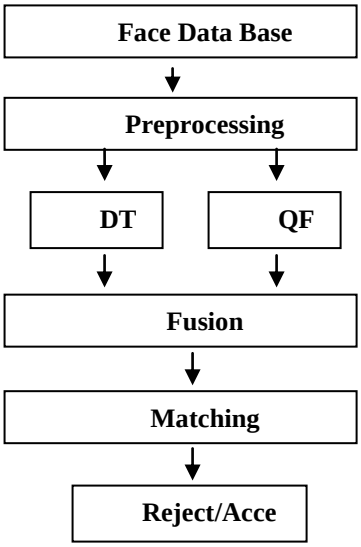

Figure 1: Block Diagram for Face Recognition

Proposed work incorporates multi scale resolution methods for obtaining face image characteristics. System independently includes DTCWT technique to obtain one set of features coefficients. The DTCWT of x (n) signal is created with two significantly sampled DWT's for the same data in parallel. The Filter bank of DTCWT is described in Figure 2.

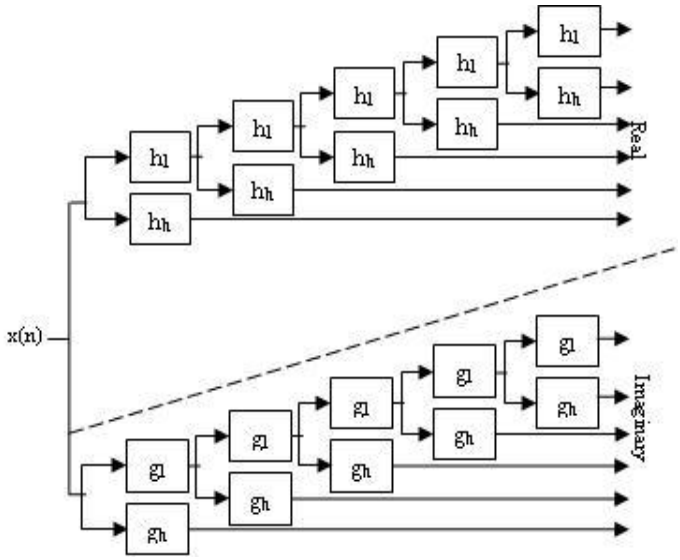

Figure 2: Filter Bank of Dual-Tree Complex Wavelet Transform

The first tree generates the real information whereas second tree produce the illusionary information of transform respectively. The high pass pair (h_h) and low pass pair (h_l) information is obtained from the real part tree and the complex coefficients of low pass (h_l) & high pass (h_h) pair of information is obtained from the imaginary part of DTCWT and also it gives the total of six bands with different angles such as ±15, ±45, and ±75. This sub band information is considered as features. The complex wavelet and complex scaling function are outlined by the following equations:

$\Psi(t)= \Psi h(t)+j\ \Psi g(t)$

$\phi(t) = \phi h(t)+j\phi g(t)$

2D complex separable wavelets and 2d complex scaling functions are described as:

$\Psi 1(p, q) = \phi(p)\ \Psi(q)$

$\Psi 2(p, q) = \Psi(p)\ \phi(q)$

$\Psi 3(p, q) = \Psi(p)\ \Psi(q)$

$\phi\ (p, q) = \phi\ (p)+ \phi\ (q)$

The final features are achieved by integrating QFT features with DTCWT features using arithmetic operation. In order to classify the test feature with trained set of features Euclidean distance (ED) classifier is used. Samples of L-spec data set of a person with different poses are shown in Figure 3.

Here QFT uses symmetry properties of cosine and sine functions to derive an efficient algorithm.

$\mathrm{COS}(2\pi(N-n)\ K)/\ N = \mathrm{COS}(2\pi nK)/\ N$

$\mathrm{SIN}(2\pi(N-n)\ K)/\ N = -\ \mathrm{SIN}(2\pi nK)/\ N$

We can write N+1 point DCT as:

$$X_{DCT}(K) = \sum_{n=0}^{N} x(n)\ \mathrm{COS}(\pi nK)/N$$

K=0,1…. N

Similarly, a N -1 point DST is written as:

$$X_{DST}(K) = \sum_{n=1}^{N=-1} x(n)\ \mathrm{SIN}(\pi nK)/N$$

K=0,1…. N

Where recursive operation occurred on all calculations of DST & DCT are combined to get QFT (Quick Fourier Transform) which is well suited for operation real data with reduced computation time.

Figure 3: L-Spacek's samples of a same person

The figure 4 shows the proposed model for speech recognition.

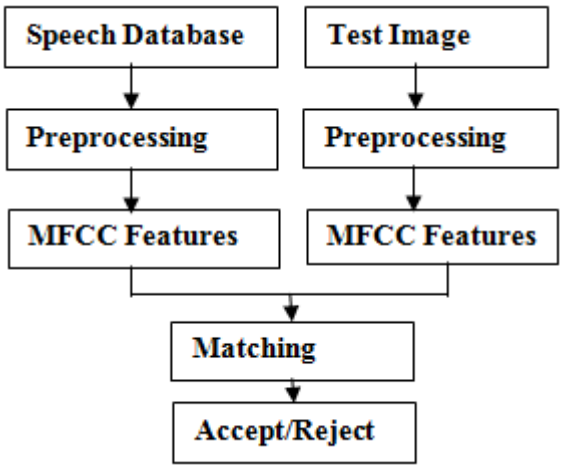

Figure 4: Block Diagram of Speech Recognition

In speech recognition voice consists of more information, we have to identify which person is speaking by extracting person's voice characteristics. In preprocessing speech signal is converted into digital representation, because the signal of speech differs with time; Although we evaluate it in time between 5 milliseconds and 100ms its characteristics are relatively unchanged. We will observe the difference in the speech signal, after 0.2 seconds or more than that, hence the better method to run audio signal is shot term spectral analysis.

MFCC (Mel-Frequency Cepstral Coefficients)

The Mel-Frequency Cepstral Coefficients method is on the basis of human hearing behavior that will not recognize frequencies greater than 1 KHz [18]. The ear of human can able to distinguish different frequencies. The MEL scale is used to express the signal and are centered on observation of pitches measured by observers at regularly spaced intervals. This scale makes use of a filter based on logarithmic spacing above 1000Hz and linearly spaced the frequencies below 1000 Hz.

Farming

Framing is the process of segmentation between the ranges of 20ms to 40ms. The voice signal is split into N sample frames and distinguished by M adjacent frames that is M<N by taking the N=256 and M=100 values. For obtaining limited length, hamming window is used with different values of N=128 and 256, the values of M= 50 and 100 and the combination of M=100 and N=256 provide better performance and FFT is implemented for converting time domain to frequency domain of N samples in every frame.

Mel Filter Bank Processing

The MFB technique is used for obtaining linear scale from the wide range of FFT spectrum of the voice signal. Figure 5 describes the filter bank.

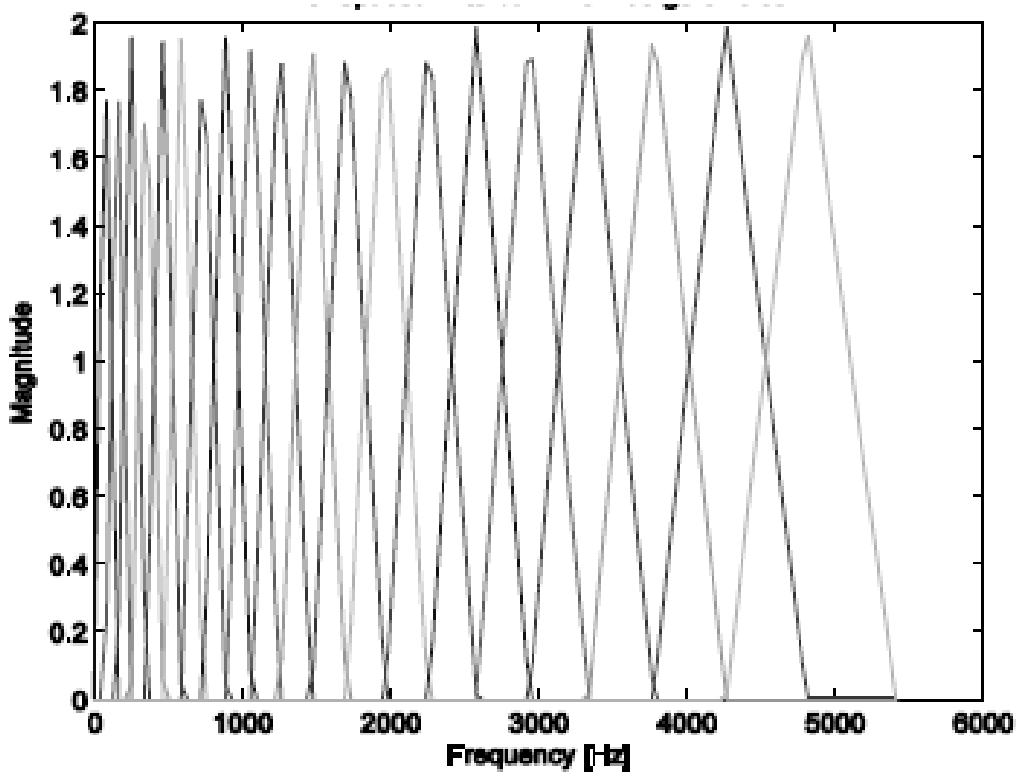

Figure 5: Graphical overview of Filter Bank

Filters are used to analyze a weighed number of spectral components, and to obtain filter output Mel scale process is used. The filter response gives the magnitude of frequency in the form of triangle and seems to be equal to unity at center frequency & decreases linearly to zero of the 2 adjoining filters at the center frequency and it gives the output. The yield is the total of filter's spectral components, which is calculated for the given frequency using the following equation in Hertz.

$$F(Mel)=[2595*\log 10[1+ f\ ]\ 700]$$

Feature Matching

From the available feature matching techniques like, HMM, DTW, & Vector Quantization [20], the Vector Qualification technique is used in the proposed work because ease to implement with better accuracy.

Distance measure

The unknown speaker's voice is characterized by a feature vector sequence of (y_1, y_2 …. yi), after it is compared with the codebooks from the database. To recognize the unknown speaker by comparing the distance of two vectors, when the distortion distance is low the person is recognizing as a known person using Euclidean distance.

To carry out the work, we can also create the speech database which contains the first 20 persons containing 6 speech signals per person.

Training

For training, from the Space Accent Achieve dataset first 20 persons 6 speech signals per person are used; hence total signals for training used are 120 signals.

For FRR and TSR calculation

For FRR calculation, seventh signal is considered from first 20 persons (Inside Database).

For FAR calculation

For FAR calculation, seventh signal is considered from signals from 21-30 persons (Outside Database).

Matching

The matching is carried out separately for face images and speech signal images using Euclidean distance. When the distance between the corresponding feature vectors based on the minimum score of the two faces and speech images being matched corresponds to the best alignment. If the Euclidean distance between two feature vectors is less than a threshold value, then the decision that "the two images are matched and these images are come from the same person and otherwise a decision that, the two images are not matched and these images are come from different person.

By using ED for matching, the features of database images the test image compared and it's calculated with the help of below equation.

$$Euclidean\ Distance = \sqrt{\sum_{i=1}^{N}(q_i - p_i)^2}$$

Where p_i = Feature values Database image.

q_i = Features value Test image.

4. Results Analysis

The L-spacek face databases and Speech Accent Archive datasets are considered to evaluate the performance of the proposed work.

Table 1: Face dataset experimented

Threshold	FAR	FRR	TSR
8.2	0	100	0
8.5	0	100	0
9	0	100	0
9.2	0	92.50	7.50
9.5	5	90	10
9.8	5	70	30
10	5	60	40
10.2	5	45	55
10.5	15	30	70
10.8	30	17.50	82.50
11	45	12.50	87.50
11.2	65	7.50	92.50
11.5	80	2.50	96.50
11.8	80	2.50	97.40
12	80	2.50	97.40
12.2	85	2.50	97.40

The value of %TSR, %FRR, and %FAR are exemplified in Table 1. It is observed that by varying the value of threshold from 0.1 to 2.3 the False Rejection Rate values reduce from 100% to 1.16% and Total Success Rate increase from zero to 98.83%. The value of %False Acceptance Rate is zero till threshold value is 0.3 and FAR value reaches to 98.52% when the threshold value reaches to 2.3.

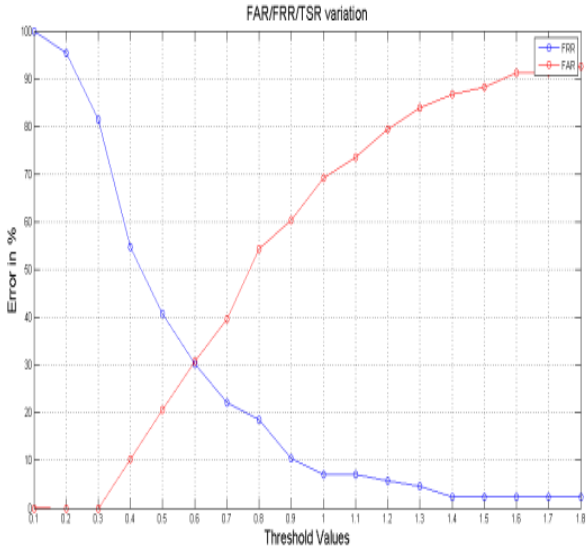

Figure 6: FRR and FAR Plot vs. L-Spacek database threshold.

Result Analysis for speech

The table 2 demonstrates the values of %TSR, %FRR, and %FAR, for speech data. The FAR value varies from zero to 85 percent by varying the value of threshold from 8.2 to 12.2.and TSR becomes 97.5%. The value of FRR is 100% till threshold value is 9 and it reduces to 2.5 when the threshold value reaches to 12.2.

Table 2: Speech dataset experimented

Threshold	FAR	FRR	TSR
0.1	0	100	0
0.2	0	95.34	4.65
0.3	0	81.39	18.60
0.4	10.29	54.65	45.34
0.5	20.58	40.69	59.30
0.6	30.88	30.23	69.76
0.7	39.70	22.09	77.90
0.8	54.41	18.60	81.39
0.9	60.29	10.46	89.53
1	69.11	6.976	93.02
1.1	73.52	6.976	93.02
1.2	79.41	5.813	94.18
1.3	83.82	4.651	95.34
1.4	86.76	2.325	97.67
1.5	88.23	2.325	97.67
1.6	91.17	2.325	97.67
1.7	91.17	2.325	97.67
1.8	92.64	2.325	97.67
1.9	92.64	1.1627	98.73

2	95.58	1.1627	98.73
2.1	95.58	1.1627	98.73
2.2	98.48	1.1627	98.73
2.3	98.48	1.1627	98.23

Figure 7: FRR and FAR Plot vs. limit for Space Accent Achieve database.

The graph of FRR and FAR is shown in Figure 6 and 7 with different threshold values at which FAR and FRR intersects.

5. Conclusion

In the proposed work face identification using DTCWT has been used effectively for L-Spacek database. The pre-processing is accomplished on face image for obtaining uniform size foe all the images and Duel-Tree Complex Wavelet Transform is used in the resized image of faces for obtaining DTCWT features & these characteristics are considered as the final ones. The Euclidean Distance is adapted for matching. We can observe that Total Success Rate is 98.83% for face database and 97.50% for speech database

References

Marcos Faundez-Zanuy, “Biometric Security Technology,” Encyclopedia of Artificial Intelligence, Vol. 1, pp. 262–264, Jan. 2008.

Tadi Chandrasekhar and Ch. Sumanthkumar, “Face Recognition System using Adaptive Neuro-Fuzzy Inference System,” International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques, pp. 448-455, Dec 2017.

Xiang-Yu Li and Zhen-Xian Lin, “Face Recognition Based on HOG and Fast PCA Algorithm,” Proceedings of the Fourth Euro-China Conference on Intelligent Data Analysis and Applications Advances in Intelligent Systems and Computing, Springer, Cham., vol.682, pp. 10-21, Sept 2017.

Ningning Zhou, •A. G. Constantinides, Guofang Huang and Shaobai Zhang, “Face recognition based on an improved center symmetric local binary pattern,” Neural Computing and Applications, 30, pp. 3791–3797, Dec 2018.

Ravi J, Saleem S Tevaramani and K B Raja, “Face Recognition using DT-CWT and LBP Features,” International Conference on Computing, Communication and Applications, pp. 1-6, 2012.

SaiyedUmer, Bibhas Chandra Dhara and Bhabatosh Chanda, “Face recognition using fusion of feature learning techniques,” Measurement, Vol. 146, pp. 43-54, Nov 2019.

Lijian Zhou, Hui Wang, Wanquan Liu and Zhe-Ming Lu, “Face feature extraction and recognition via local binary pattern and two-dimensional locality preserving projection,” Multimedia Tools and Applications, 78,

Tong Xiao, Jingbing Li1, Jing Liu1, Jieren Chengand Uzair Aslam Bhatti, “A Robust Algorithm of Encrypted Face Recognition Based on DWT-DCT and Tent Map,” Cloud Computing and Security,Lecture Notes in Computer Science,Springer, Cham., vol. 11064, pp. 508-518, 2018.

Eyad I. Abbas and Mohammed E. Safi, “Effect of Wavelet Decomposition on Database Size Reduction for Face Recognition Rate,” International Conference on Advanced Science and Engineering, pp. 35-39, 2019.

K. PunnamChandar and T. Satyasavithri, "DT- CWT Sub-band Partitioning for Face Recognition," International Conference on Industrial and Information Systems, pp. 1-5, 2014.

Hua Wang, DingSheng Zhang and ZhongHua Miao, "Fusion of LDB and HOG for Face Recognition," Chinese Control Conference pp. 9192-9196, 2018.

Chunling Tang, Min Li, "Speech Recognition in High Noise Environment," Foundation Environmental Protection & Research-FEPR, pp. 1561-1565, 2019.

Lucas Debatin, Aluizio Haendchen Filho and Rudimar L. S. Dazzi1, "Offline Speech Recognition Development," International Conference on Enterprise Information Systems, pp. 551-558.

R. Thiruvengatanadhan, developed an algorithm Speech Recognition using AANN, International Journal of Innovations in Engineering and Technology, vol. 12, pp. 72-75, 2019.

Ritesh A. Magre, Ajit S. Ghodke, Robust feature extraction for visual speech and speaker recognition, International Journal for Research in Engineering Application & Management, pp. 33-35, 2019.

Mehryar Mohri, Fernando Pereira and Michael Riley, "Speech Recognition with weighted Finite-State Transducers," Springer Handbook on Speech Processing and Speech Communication, pp.1-31.

Ashok Kumar, Vikas Mittal, "Speech Recognition: A Complete Perspective," International Journal of Recent Technology and Engineering, vol. 7, pp. 78-83, 2019.

T. R. Jayanthi Kumari and H. S. Jayanna, "Limited Data Speaker Verification: Fusion of Features," International Journal of Electrical and Computer Engineering, VOL. 7, PP. 3344-3357. 2017.

P. Krishnamoorthy, H.S. Jayanna, S.R.M. Prasanna, "Speaker recognition under limited data condition by noise addition, Elsevier Expert Systems with Applications, vol. 38, pp. 13487-13490, 2011.

Hector Perez-Meana and Enrique Escamilla-Hernandez, "Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques", IEEE International Conference on Electrical Communications and Computers. PP. 248-251. 2012.