An Automated Fraud Detection of Hall Ticket in an Offline Examination System using ANN Classifier

Shridevi Soma
Computer Science and Engineering
PDACE, Kalaburgi

Vidyashree S.A.
Computer Science and Engineering
PDACE, Kalaburgi

ABSTRACT

Document image analysis is a major research area in pattern recognition system for identification image objects within a document and their classification. Face recognition is most widely used and efficient system in biometric. Existing methods worked on these areas of document image analysis and face recognition separately. In this paper, an attempt is made to develop a novel system to authenticate a person to appear exam by comparing name, seat number, and photograph images on the hall ticket . An ANN classifier is used to detect fraud in the exam hall tickets. The main aim of the proposed work is to providing security technique for offline conduction of exams using hall ticket. At a coarse scale, hall ticket identification system also comprises of five phases: Preprocessing, Segmentation, Face detection, Feature extraction and ANN classification. For feature extraction 10 shape features and 6 histogram features of object occurrence in a new tessellation of hall ticket images are used and these features are given to the ANN classifier for recognizing the fraud in hall tickets. The proposed algorithm is experimented on a data set of two different institutes, with an average accuracy of   94.12%.

General Terms

Pattern recognition, Adaboost algorithm, ANN Classifier.

Keywords

Preprocessing, Segmentation, Face detection, Feature extraction.

1. INTRODUCTION

Hall tickets are commonly used in both online and offline examination. The problem of fraud in hall ticket detection and recognition is of great interest in the document domain as it enables us to identify the ownership of the document. In the context of document image retrieval, hall tickets provide an important form of indexing that enables effective exploration of data. Given a large collection of hall tickets, detecting for a fraud hall ticket is a highly effective way of retrieving documents. The main focus of this paper is to detect fraud in the exam hall tickets which consists of textual elements like name, register number, etc and graphical elements like photo of the candidate. The system is aimed to detect unauthorized person appearing for examination i.e., by replacing the original candidate photo with an unauthorized person.

 In the past history work has been carried out on examination malpractice [1], detecting the fraud in Selection Exams by Using Knowledge Engineering Tools [2]. Prior Literature[3,4] focused on online test, using a multi model authentication technique, that helps in continues monitoring of the student. The behavior of the student is visualized by using different visualization techniques. Stephan Kovach et. al. [5] have

presented the Online Banking Fraud, and global behavior is specified to improve the fraud in the accounts that has been accessed. A. Brabazon et. al. [6] have developed online credit card fraud and Artificial Immune Systems, to calculate the effectiveness of Artificial Immune Systems (AIS) for credit card fraud detection using a large dataset obtained from an on-line retailer. Three AIS algorithms were implemented and their performance was benchmarked against a logistic regression model. Shailesh S. Dhok [7] presented credit card fraud transaction for online shopping, paying bills, it is shown that credit card fraud can be detected using Hidden Markov Model during transactions. Mohd Avesh Zubair Khan et. al. [8] system to detect the Credit Card developed a Fraud, modeled the sequence of transactions in credit card processing using an HMM and k-means clustering A. K. Jain, A. Ross, and S. Prabhakar, U. Park, Y. Tong, and A. K. Jain, R. Gross, S. Baker, I. Matthews, T. Kanade, G. Hua and A. Akbarzadeh [9, 10, 11, 12] focused on face Recognition system. All these studies reveals that face identification and its texture based approaches are available.

From the above literature survey it can be observed the need of switching from manual to electronic fraud detection system for conducting many online/offline examination. As in Indian education system majority of the exams carried out in offline with a hall ticket/ Admit card as an authentication media, this motivated to develop a system to detect fraud in the offline examination. Hence the proposed system.

The main objective of this work is to detect whether the person is authorized to write exams by using his/her hall ticket, to achieve this it is planned to develop a methodology to segment candidate name, seat number and photo which is located on the hall ticket image and trained for authorized person and then these segmented sub-images are matched with complete hall ticket image under test to detect the fraud.

An ANN classifier is used for similarity measure between trained and test features. Adaboost algorithm is used for face recognition and applied on only segmented photo image of the hall ticket. When scanned paper documents are analyzed to produce structural electronic representations or for the purpose of sorting by corporate identity, hall ticket recognition becomes an important component. Successful recognition of hall tickets facilitates automatic source classification of document images and may also be used to determine how best to process the information contained within these particular documents. Sample hall ticket is presented in Fig 1 below. In the proposed work hall ticket images are given as a query image.
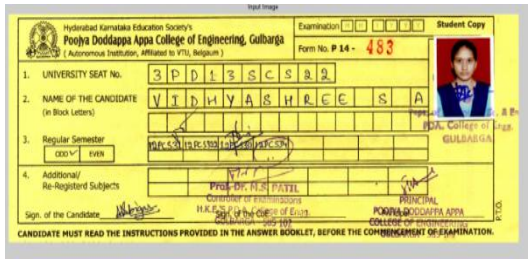
Fig 1: Sample hall ticket image

The rest of this paper is organized as follows. Next Section deals with the phases of proposed work such as image acquisition and database creation, preprocessing, segmentation, face detection, feature extraction, ANN classification. The Experimental Result is provided in Section II and Section III summarizes the paper in the form of Conclusion and Future scope.
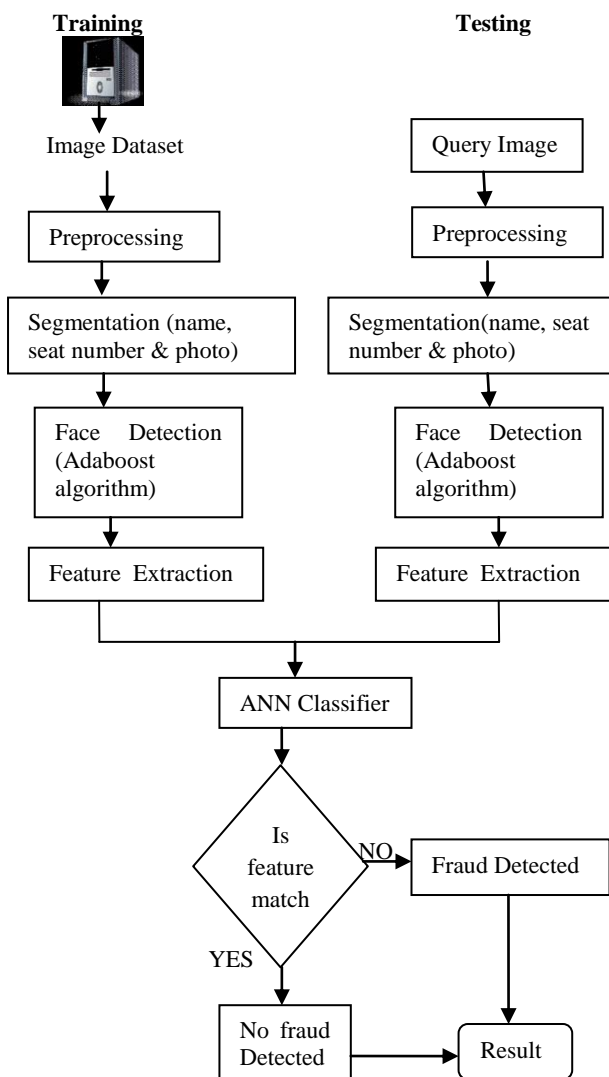

Fig.2: Block diagram of automatic hall ticket recognition

The block diagram of automatic hall ticket recognition includes the combined training and testing dataset using the phases such as Image acquisition, Preprocessing, Segmentation, Feature extraction, Face detection, ANN classifier. As shown in fig 2. The following is the algorithm for hall ticket recognition.

Algorithm: Hall Ticket Recognition

Input : Colored images of hall tickets.
Output : Authentication flag(Authorized or Unauthorized person)
Method : ANN Classifier

Train phase:

Start
    Step 1 : Preprocessing Input images
    Step 2 : Segment the images by using
        morphological operations
    Step 3 : Extract features using shape features and
        histogram features
    Step 4 : Detection of face using Ada boosting
        algorithm
    Step 5: ANN Classification
End

Testing phase:

Start
    Step 1 : Preprocessing Input image
    Step 2 : Segment the image by using morphological
        operations
    Step 3 : Extract features using shape features and
        histogram features
    Step 4 : Detection of face using Ada boosting
        algorithm
    Step 5 : ANN classifier for feature match.
        If Feature matches test image results as
        authorized person otherwise
        unauthorized person
End

1.1 Image Acquisition & Database creation

The images are acquired by scanning. Scanning is a way of changing the exam hall ticket document into digital format. The exam hall ticket images are scanned by the HP Photo smart C4388 series scanner. The images are scanned at 300 dpi resolution which generates an image of 3510 X 2550 pixels and are added to train data set.

1.2 Preprocessing

In image processing, main purpose of preprocessing stage is to enhance the image in ways that raise the opportunity for success of the other processes. Preprocessing enhances the quality of the input image. In this work, the system of Preprocessing includes the resizing and binarization.

Resizing is used to change the size of an image or scale an image. Normalizing the images by bringing them to common resolution by converting all sizes of images   200 X 200 pixels.

The color and gray scale images are converted to binary format. This process of conversion is known as binarization. The process of binarization is shown in the fig 3.

Fig.3 : Process of Binarization

The hall ticket image is in color image form, it is converted to gray scale image and this gray scale image is converted into the binary image for further processing.

1.3 Segmentation

Segmentation subdivides an image into its constituent regions or objects. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. In the proposed work the segmentation is performed using morphological operations. After binarization, document images are usually filtered to reduce noise. For documents, more specific filters can be designed to take advantage of the known characteristics of the text and graph components. A hall ticket to be scanned can itself be contaminated with dust or spots etc which constitute noise. Scanning itself can introduce some amount of noise. Noise is also due to the degeneration, ageing, photocopying or during data capture. In order to make it suitable for further processing, a scanned document image is to be freed from any existing noise.

The hall ticket image contains name , seat number, and photo or face and preprocessed to remove noise. For this noise free hall ticket image, segmentation is carried out by performing fundamental morphological operations. Dilation is performed twice by using a line structuring element. The contents of the document image are in the foreground, this foreground content thickens when dilation is performed. The hall ticket is dilated horizontally as well as vertically controlled by a line structuring element. Finally, the image is segmented into its individual components. The segmented hall ticket for name, seat number and photo from all the images are collected and stored in a database.

Dilation **:** Dilation is an operation which grows or thickens objects in binary image. The extent of thickening is controlled by a flat line structuring element. Dilation is a process where ON (white-color pixel) valued layers are added to boundaries to increase their size, which in turn reduces the size of the OFF regions(black-color pixel). The fig 4, shows the process of dilation.
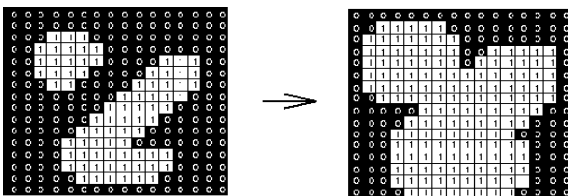

Fig 4: Process of dilation

Structuring Element **:** The structuring element consists of a pattern specified as the coordinates of a number of discrete points relative to some origin. The origin is marked by a ring around that point. The fig 5, shows the structuring elements.
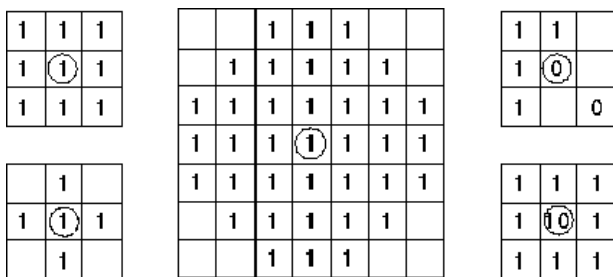

Fig 5: Structuring elements

Bounding Box: A bounding box for an object is just a rectangular box in three-dimensional space, with sides parallel to the coordinate planes, that contains (or surrounds) the object. This illustration below shows a two-dimensional box surrounding a curved object.
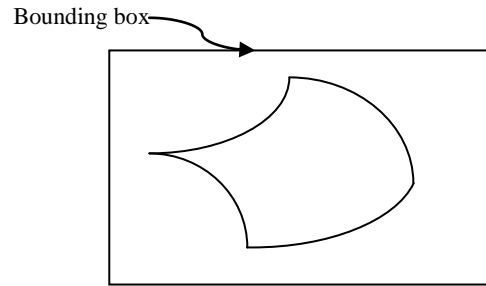
Bounding box

Fig 6: Bounding box

1.4 Face detection

Face detection is a difficult task in image processing which has wide range of applications. The existing methods for face detection can be divided into two types, image based methods and feature based methods. We have developed an intermediate system, using a boosting algorithm to train a classifier which is capable of processing images rapidly while having high detection rates. It is an aggressive learning algorithm which produces a strong classifier. Simple classifier contains simple rectangular wavelets which are reminiscent of the Haar basis. Their simplicity and a new image representation called integral image allow a very quick computing of these haar-like features. There are three kinds of Haar-like features. 1) two-rectangle features, 2) three-rectangle features, and 3) four-rectangle features. The value of a two-rectangle feature is the difference between the sum of the pixels within two rectangular regions. The regions have the same size and shape and are horizontally or vertically adjacent. A three-rectangle feature computes the sum within two outside rectangles subtracted from the sum in a center rectangle. Finally, a four-rectangle feature computes the difference between diagonal pairs of rectangles. the following fig 7, shows the three kinds of rectangle Harr-like features as follows:
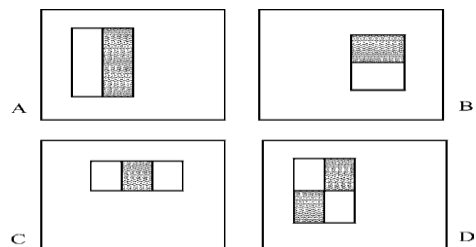

Fig 7: Rectangle Haar-like features.

A and B are the two-rectangle features, C is the three-rectangle features, and D is the four-rectangle features.

The input of this first stage is a pre-processed image and the output of the neural network is a real value between -1 and +1. The original image is decomposed in a pyramid of images (by simple sub-sampling) in order to detect faces larger than the basic detector size. There are three types of hidden units to represent local features that represent well faces. This first stage yields good detection rates, but it remains still an insufficient false positive rate.

The detector consists in scanning an image at different scales. Each image is tested by a classifier made of several stages. If it is clearly not a face, it will be rejected while more specific classifier will classify it if it is more difficult to discriminate.

Then a structure in cascade is introduced in order to reject quickly the easy to classify background regions and focus on face image. The detection of faces in input images is

proceeded using a scanning window at different scales which permits to detect faces of varying size without re-sampling the original image.

The boosting techniques improve the performances of base classifiers by re-weighting the training examples. The learning using Boosting is the main contribution of this face detection.

In this work, haarcascade - adaboost algorithm is used. The xml file for frontal face is loaded, which removes the unwanted background regions and it will focus on the face image. Then, finding the area of detected face object and find the maximum area of the calculated area. Next, draw the rectangle over the detected face image and hold on face location. The face image is cropped in rectangular form to detect the correct face image.

1.5. Feature Extraction

The feature is a function of one or more measurements, each of which specifies some quantifiable property of an object, and is computed such that it quantifies some significant characteristics of the object. Features such as shape, histogram, texture, color, etc. are used to describe the content of the image.

In shape feature method, first preprocessing done and then measurable properties of image regions are obtained. The 10 shape features mentioned below are used in this work and find the mean of all these values are stored as feature vector.

Area **:** It is defined as actual number of pixels in the region.

Perimeter **:** It is defined as the distance around the boundary of the region.

Form factor **:** The pattern of scattering white pixels in an image within the bounding box.

Major Axis **:** It is defined as the length of the major axis of the ellipse that has same normalized second central moments as the region.

Minor Axis **:** It is defined as the length of the minor axis of the ellipse that has same normalized second central moments as the region.

$$\text{Roundness} = \frac{4 \times Area}{\pi \times Majoraxis^2} \qquad [1]$$

Compactness **:** Compactness is an indication of solidness and convexity.

Density **:** Density is defined as the area of white pixels within the bounding box. It is the ratio between area of white pixels within the bounding box and the area of bounding box(BB) which is given by:

$$Density = \frac{Area\ of\ white\ pixel\ within\ the\ BB}{Area\ of\ BB} \qquad [2]$$

Black Pixel Of Each Line **:** This feature is defined as the number of black pixels in each line (BPEL) to the width of the bounding box. This is given by:

$$BP = \sum \frac{Number\ of\ black\ pixels\ of\ each\ line}{Width\ of\ bounding\ box} \qquad [3]$$

Vertical Projection Variance **:** The vertical projection of black pixels within the bounding box is found and then the variance of only the vertical coordinates of the vertical projection profile is computed.

In histogram feature method, first input parameters are checked and find the histogram for all scaled gray level from 1 to N and then calculate its statistics i.e. mean, variance, skewness, kurtosis, energy, entropy. Then, obtain the approximate probability density of occurrence of its intensity levels and find the utility functions, called as feature vector.

$$P_n = \frac{Number\ of\ pixels\ with\ intensity\ n}{Total\ number\ of\ pixels} \qquad [4]$$

Where n = 0,1,…L-1

Mean **:** It is nothing but the average i.e. adding up all the numbers. Mean is defines as:

$$\bar{X} = \frac{\sum X}{N} \qquad [5]$$

Where $\sum$ represents the summation
X, represents pixel
N, represents number of pixels.

Variance **:** The variance of a data set is the arithmetic average of the squared differences between the values and the mean. Thus, the variance of a frequency distribution is given by

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} \qquad [6]$$

Where s^2 = Variance, Σ = Summation, which means the sum of every term in the equation after the summation sign, x_i = Sample observation. This represents every term in the set, $\bar{x}$ = mean, n = size.

Skewness **:** Skewness is a measure of the asymmetry of the data around the sample mean. If skewness is negative, the data are spread out more to the left of the mean than to the right. If skewness is positive, the data are spread out more to the right. The skewness of the normal distribution is zero.

$$s = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^n (X_i - X_{avg})^3}{S^3} \qquad [7]$$

Where s is the skewness, n is the size.

Energy: The energy is defined as follows:

$$Energy = \sum P_n \qquad [8]$$

Entropy: Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. Entropy is defined as

$$H(P) = -\sum_{i;n} p(s_i) * \log(p(s_i)) \qquad [9]$$

Where H is the entropy

Kurtosis: Kurtosis is a measure of how outlier-prone a distribution is. The kurtosis of the normal distribution is 3. Distributions that are more outlier-prone than the normal distribution have kurtosis greater than 3; distributions that are less outlier-prone have kurtosis less than 3.

$$K = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^n (X_i - X_{avg})^4}{S^4} \qquad [10]$$

Where K is the kurtosis, n is the size

1.6. Train the ANN

In this phase, the candidate name, seat number and photo are trained. The back-propagation technique is akin to supervised learning as the network is trained with the expected reply/replies. Each iteration modifies the connection weights in order to minimize the error of the reply (expected value-estimated value). Adjustment of the weights, layer by layer, is calculated from the output layer back to the input layer. This correction is made by:

$$\Delta W_{ji} = \eta \delta_j f(a_i) \qquad [11]$$

Where Δw_{ji} is the adjustment of weight between neuron j and neuron i from the previous layer; $f(a_i)$ is the output of neuron i, η is the learning rate, and δ_j depends on the layer. For the output layer, δ_j is:

$$\delta_j = (Y_j - \hat{Y}_j)f_i^{'}(a_j) \qquad [12]$$

Where Y_j is the expected value ('observed value') and $\hat{Y}_j$ is the current output value ('estimated value') of neuron j. For the hidden layer, δ_j is:

$$\delta_j = f_j^{'}(a_j)\sum_{k=1}^{K}\delta_k W_{kj} \qquad [13]$$

Where K is the number of neurons in the next layer. The learning rate plays an important role in training. When this rate is low, the convergence of the weight to an optimum is very slow, when the rate is too high, the network can oscillate. To reduce these problems, a momentum term α is used and ΔW_{ji} becomes:

$$\Delta W_{ji} = \eta\delta_j f(a_i) + \alpha\Delta W_{ji}^{Prev} \qquad [14]$$

Where ΔW_{ji} denotes the correction in the previous iteration.

The training, performed on a representative data set, runs until the mean squared of errors (MSE) is minimized:

$$MSE = \frac{1}{2}\sum_{p=1}^{P}\sum_{j=1}^{N}(Y_{pj} - \hat{Y}_{pj})^2 \qquad [15]$$

Where Y_{pj} is expected output value, $\hat{Y}_{pj}$ is the estimated value by the network, $j = 1 . . . N$ is the number of records and $p = 1 . . . P$ is the number of neurons in the output layer.

The structure of the network, the number of records in the data set and the number of iterations that determine the training duration. In our study, for a 51 records characterized by 10 input variables and two output variables, and only one hidden layer with 20 neurons, 1000 iterations last about 5 minutes with an Intel 486 DX2–66 processor.

The Fig 8, shows the sample of training dataset as follows:

Fig 8: Training Dataset

1.7 Testing

After training, the performance of the network has to be tested. As in discriminant analysis, a first indication is given by the percentage of correct classifications of the training set records. Nevertheless, the performance of the network with a test set (set of similar data unused during training) is more relevant.

In the test step, the input data are inserted into the network and the desired values are compared to the network's output values. The authenticated or unauthenticated of the results thus give an indication of the performance of the trained network.

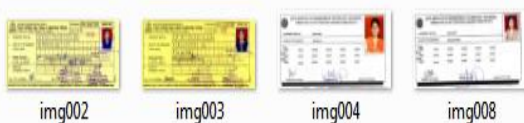The fig 9, shows the testing dataset from a two different institutes :

Fig 9: Testing dataset

1.8 Classification

Classification is the task of arranging the data in group and classes according to resemblances and similarities. Classification algorithms divided into two phases of processing: training and testing. In the initial training phase, image features are isolated for the characteristic of input image properties and, based on these, a unique description of each classification category, known as training class, which is created. In the next testing phase, these feature partitions are used to classify image features. The various classifiers are Bayesian learning, Artificial neural network(ANN), Support vector machine(SVM), and so on. In this paper, ANN Classifier is used. Artificial neural network is used in the automatic detection of fraud in hall tickets. Artificial neural network is chosen as a classification tool due to its well known technique for many real time applications. The training and testing are among the important steps in developing an accurate process model using ANN. The hall ticket image is presented to a feed-forward network with one hidden layer. The unit in the hidden layer are locally connected to the units in the input layer, forming a set of local feature maps. The output unit is interpreted as revised probability of the input pattern's belonging to a particular class.

Fig 10: Classification of input image

1.9. Classification Result

The output categories are ordered according to the activation levels and passed to the post processing stage. In this stage, the candidate photo is matched with the candidate name and seat number. If these three i.e., candidate name, seat number, and photo are matched then system gives the result as person is authenticated or if these three are mismatched then system gives the result as person is unauthenticated. The classified result shown in the fig 11.
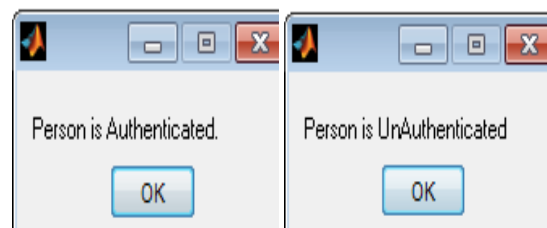

Fig 11 : Classified Result

2. RESULTS AND DISCUSSION

The feature.dat and outtype.dat file is generated after completion of the training dataset. The images of train dataset are stored in this file. There are 45 images and 78 images of the hall tickets for Inst 1 and Inst 2 computed for feature.dat and outtype.dat

The testfeature.dat file is generated after completion of testing dataset. The features of the test image or the query image are stored in this file. For testing 17 images and 26 images are considered.

Table 1 : Identification rate

Types of hall ticket images	Training dataset	Segmented images (photo, name and seat number)	Number of hall tickets identified		Identification rate
			Correctly	Incorrectly	
Inst1	17	45	15	2	88.24%
Inst2	26	78	26	0	100%
			Accuracy rate		**94.12%**

The above classification percentage table shows the classification rate of Inst 1 and Inst 2. For Inst 1 both dataset contains the two folders, in training dataset first folder contains the 36images and second folder contains the 9images, testing contains 17 images. Out of these 17images, 14images are authenticated in that 2images are the failures and 3images are unauthenticated. Hence, the identification rate is 88.24%. For Inst 2, training contains 78 images and 26images for testing. Out of 26 images, 25 images are authenticated and 1 is unauthenticated. Hence, the identification rate is 100%. Hence the accuracy rate is 94.12%.

The Fig 12 shows the graph of performance for testing the two different institutes
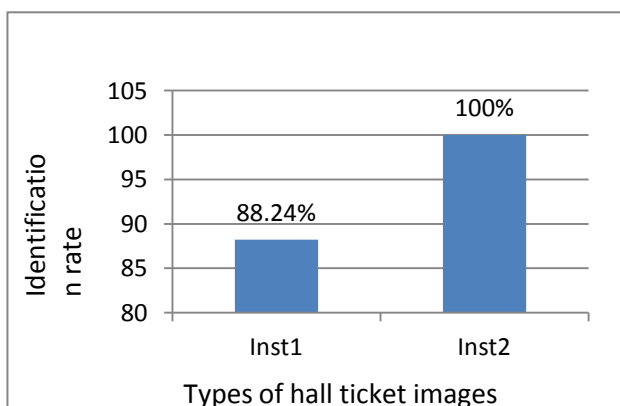
Fig 12: Performance graph for hall ticket recognition by ANN Classifier

3. CONCLUSION AND FUTURE SCOPE

The detection of fraud in hall tickets is a challenging task for the purpose of classifying the distinct image entities in an automated approach. The hall ticket images are used in the proposed system for off-line exams to provide security based on their entities. The hall ticket images are processed and segmented using morphological operation. Each segment is labeled using connected component labeling technique. The features such as area, form vector, roundness, compactness, density, etc are extracted. ANN classifier is used to detect the fraud in hall ticket images i.e., if candidate name, seat number is mismatched with candidate photo then the system displays person is unauthenticated. The system is developed for the offline examination system to detect the fraudulent hall ticket.

The future scope this work is to combined with online biometric or face detection system to enhance security level.

4. REFERENCES

[1] Adams O. U. Onuka1 , Esther O. Durowoju2 "Stakeholders' Role in Curbing Examination Malpractice in Nigeria". International Journal of Economy, Management and social sciences, 2(6) June 2013 pp:342-348

[2] Marcus de Melo Braga & Mario Antonio RibeiroDantas "Fraud Detection in Selection Exams Using Knowledge Engineering Tools". Universidade Federal de santacatarina(UFSC) Florianopolis, Brazil.

[3] HarishBabu. Kalidasu, B.PrasannaKumar, Haripriya.P "A Fraud Detection based Online Test and Behavior Identification Implementing Visualization Techniques". IJCSET, August 2012,vol 2,issue 8 pp:1338-1344,ISSN:2231-0711

[4] Sri Anusha.N1, Sai Soujanya.T2 DrS.Vasavi3 "Study on Techniques for Providing Enhanced Security During Online Exams". International Journal of Engineering Inventions, ISSN:2278-7461,vol 1,issue 1(Aug 2012) pp:32-37

[5] Stephan Kovach and Wilson Vicente Ruggiero "Online Banking Fraud Detection Based on Local and Global Behavior".ICDS 2011:The fifth International Conference On Society pp:166-171

[6] A. Brabazon, J. Cahill, P. Keenan, D. Walsh "Identifying Online Credit Card Fraud using Artificial Immune Systems". UCD Business school,UniversitycollegDublin, Dublin 4, Ireland.

[7] Shailesh S. Dhok "Credit Card Fraud Detection Using Hidden Markov Model". International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012 pp:88-92

[8] MohdAveshZubair Khan1, Jabir Daud Pathan2, Ali HaiderEkbal Ahmed3 "Credit Card Fraud Detection System Using Hidden Markov Model and K-Clustering". International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2014 pp: 5458-5461

[9] A. K. Jain, A. Ross, and S. Prabhakar. "An introduction to biometric recognition," CSVT, vol. 14, no. 1, pp. 4–20, 2004.

[10] U. Park, Y. Tong, and A. K. Jain, "Face recognition with temporal invariance: A 3d aging model," in IEEE International Conference on Automatic Face and Gesture Recognition, 2008.

[11] R. Gross, S. Baker, I. Matthews, and T. Kanade, "Face recognition across pose and illumination," in Handbook of Face Recognition, S. Z. Li and A. K. Jain, Eds. Singer, 2005, pp. 193–216.

[12] G. Hua and A. Akbarzadeh, "A robust elastic and partial matching metric for face recognition," in Proc. ICCV, 2009.