Performance Evalution of Multistage Offline Marathi Script Recognition System

Vijaya Rahul Pawar
Bharati Vidyapeeth Deemed University College of Engg.
Pune -Satara Road,
Pune, Maharashtra, India

Arun Gaikwad, Ph.D
University Of Pune
Zeal Education Society's DESCOE&R, Narhe,
Pune, Maharashtra, India

ABSTRACT
Handwriting is the most effective way by which civilized people speaks. Devanagari is the basic Script widely used all over India. Many Indian languages like Hindi, Marathi, Rajasthani are based on Devanagari Script. Devanagari Scripts Hindi language is the third common language used all over the word. In the proposed work an artificial neural network based classifier and statistical and structural method based feature extraction approach is used for the recognition of the script. Optical isolated Marathi Characters are taken as an input image from the scanner. An input image is preprocessed and segmented. Features are extracted in terms of various structural and statistical features like End points, middle bar, loop, end bar, aspect ratio etc. Feature vector is applied to Self organizing map (SOM) which is one of the classifier of an artificial neural Network.SOM is trained for such 5000 different characters collected from 500 persons. The characters are classified into three different classes. The proposed classifier attains 93% accuracy.

General Terms
Classification algorithm used is Self Organizing map and Feature Extraction Technique used are Structural and Statistical feature extraction techniques..

Keywords
Image Preprocessing, Feature Extraction, Network Neighborhood, Self Organizing Map, Accuracy,

1. INTRODUCTION
Researchers are taking efforts in the recognition of handwritten script. Offline word recognition is challenging due to variety of the human writing. A person's handwriting is unique just like fingerprint. Offline handwriting script recognition is having its potential applications in bank automation, Cheque reading etc. A script is defined as a graphical form of a system. Six different scripts popularly used all over the world are [1] Arabic, Roman, Cyrillic, Han, Hebrew, and Devanagari .Different script may follow the same writing method. Script is written from left to right (few exceptions are like Gujarati,Oriya).Bhojpuri,Gujari,Pahari (Garhwali,Kumaoni),Konkani,Magahi,Maithili,Marwari,Bhili, Newari,Santhali,Tharu and sometimes Sindhi, Dongari,Sherpa and Kashmiri. Devanagari script is used for writing different local Indian languages and region specific variations in terms of shape of numerals specially in case of old edge people. An interesting and surprising thing in Indian handwritten document is either the word or the numeral is written in terms of English. It may have the reason like English is India's Official language. According to the survey of Indian languages conducted by Badoda Language centre there are 970 languages in which Indian people communicates. In Maharashtra there are 60 languages and sub languages out of which only two languages are having their own script(Marathi, Gond). In the present paper we are focusing on recognition of Marathi script. Because of complex patterns, large number of classes involved in basic characters , different writing styles and sizes handwritten Marathi character recognition is one of the most difficult task, It requires a large amount of time as recognition module has three stages preprocessing, feature extraction and classification. Devanagari script consist of 49 characters and 10 numerals, out of which 13 are vowels and 36 are consonants. There are 5 upper modifiers and 3 lower modifiers with side modifiers . Vowels can be written as independent letters, or by using a variety of diacritical marks which are written above, below, before or after the consonant they belong to. When vowels are written in this way they are known as modifiers and the characters so formed are called conjuncts. Sometimes two or more consonants can combine and take new shapes. These new shape clusters are known as compound characters. All the characters have a horizontal line at the upper part, known as Shirorekha or headline. In continuous handwriting, from left to right direction, the Shirorekha of one character joins with the Shirorekha of the previous or next of the same word. In this fashion, multiple characters and modified shapes in a word appear as a single connected component joined through the common Shirorekha. All these variations make the handwritten character recognition, a challenging problem. India is Many times two neighboring characters are overlapping and Many pattern recognition system are having following significant components which are shown in figure 1. Handwritten Devanagari characters are acquired as an Optical image by using scanner and converted into digital images. After pre-processing the image is applied to the segmentation . The feature extractor extracts the properties which are useful for classification. the recognition becomes difficult. The most challenging step in word recognition is segmentation of the word. Artificial Neural Network are becoming wave of the future in computing and becoming powerful class of model .Handwriting recognition using artificial neural network is most aimed at digit recognition.

1.1 Pattern Recognition system

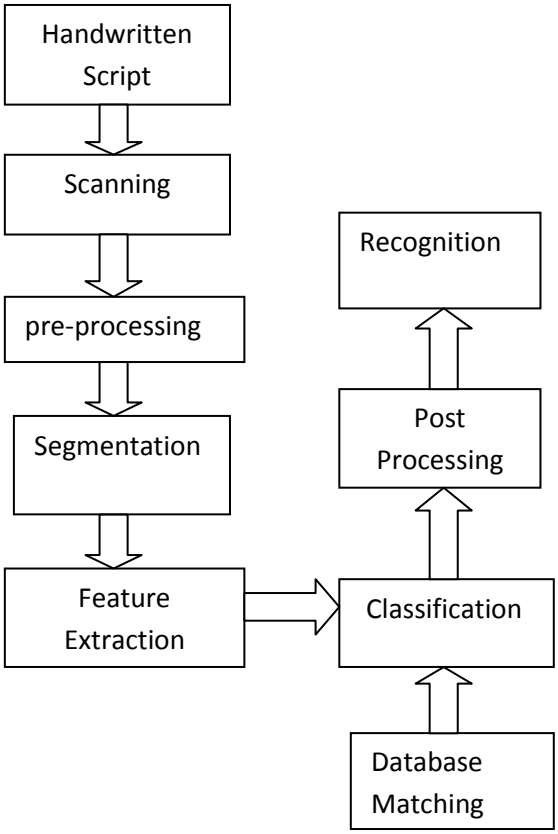


Fig 1: Pattern Recognition System

The classifier utilizes these features to assign the Input character into a class. In post processing other considerations like cost of errors are considered for the final decision.

2. PAST REVIEW

A.K.Jain [1] automatic identification of handwritten script facilities many important applications such as automatic transcription of multilingual documents and search for documents on the web containing a particular script . The increase in usage of handheld devices which accept handwritten input has created a growing demand for algorithms that can efficiently analyze and retrieve handwritten data. This paper proposes a method to classify words and lines in an online handwritten document in to one of the six major scripts: Arabic, Devanagari , Cyrillic,Han,Hebrew, or Roman. The classification is based on 11 different spatial and temporal features extracted from the strokes of the words. The proposed system attains an overall accuracy of 87.1%at world level with 5 fold cross validation on data set containing 13,379 words. The classification accuracy improves to 95% as the number of words in the test sample is increased to five and to 95.5 %for complete text lines consisting of an average of seven words.

Ujjawal Bhattacharya[2] had proposed a system that concerns the problem of isolated handwritten numeral recognition of major Indian scripts the main contribution that he has stated multistage recognition scheme and pioneering development for the database of Indian numerals. He has used three MLP classifier. If require one classifier may be added .The system has been developed for mixed numerals .He also stated that no standard database is available for Indian scripts. He has simulated the proposed recognition scheme on handwritten numeral database of three major scripts i.e. Devanagari ,Bangla etc.For Devanagari and Bangla he has created his own database but for English he has taken MNIST database he has tried the system for Devanagari English, Bangala -English as this type of manuscripts appears in various forms and mail pieces .PIN code is also written sometimes in mixed manner.

Jenowa Park Proposed offline handwritten word recognition(HWR)[9].The key ideas are actively and successfully select a subset of features for each word image which provides the minimum required classification accuracy to get a valid answer and to derive a consistent decision metric which works in multi resolution feature space and considers the interrelationship of lexicon at the same time .A recursive architecture based on interaction between flexible character classification and deductive decision making is developed. The recognition process starts from the initial coarse level using a minimum number of features, then increase the discrimination power by adding other features adaptively and recursively until the result is adapted by the decision maker .For the computational aspect of a feasible solution ,unified decision metric ,recognition confidence ,is derived from two measurements: pattern confidence, evaluations of absolute confidence using shape features and lexical confidence evaluation of the relative string dissimilarity in the lexicon He has implemented the same for the US MAIL Address component.

Reena Bajaj and Lipika Day[12] describes an efficient and reliable technique for recognition of handwritten numerals. Three different types of features have been used for classification of numerals. In this paper a new approach for recognition of handwritten Devanagari numerals using multiple neural classifiers hence tried to exploit information about stylistic variations, similarity between numerals and style invariant features. The overview of the scheme presents, the set of first type of features provide coarse shape description of the numerals and are relatively insensitive to minor changes in character shapes. But these features are not expected to remain unchanged for different writing styles.

3. RECOGNITION OF HANDWRITTENDEVANAGARI OCR SYSTEM

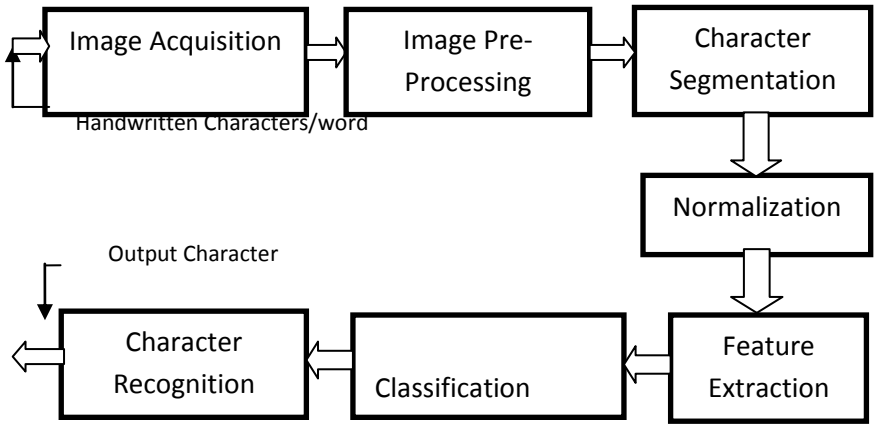


Fig.2. Recognition of handwritten Devanagari characters

The scanned Image is preprocessed . Major steps involved in recognition of handwritten characters are [11] Image Acquisition.

Image Pre-processing.
RGB To Gray conversion
Noise Removal
Segmentation.
Feature Extraction.
SOFM Based Character Classification.
Post-processing.

3.1 Image Acquisition

Handwritten Devanagari words are captured from scanner and are converted into digital images. The scanner used is 300 dpi flat bed scanner. The Devanagari words are taken from different 500 persons which are size invariant.

3.2 Pre-processing

Preprocessing relates to image preparation for the later analysis and use. The main aim of preprocessing is to make the image ready for the feature extraction. Color image can be converted into grey scale image

3.3 RGB to Gray Conversion

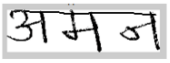
by using the formula [10]
GRAY = 0.299*R + 0.587*G + 0.114*B

3.4 Noise Removal

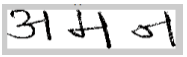
There are many methods to remove the noise from the image but Median filter with 3*3 matrix is used for noise removal .It is the statistic filter. Scanning process introduces irregularities such as ‘’speckle noise”, “salt and pepper noise” in the digital image. Median Filtering has been employed, to remove such noise and its effects, where each pixel is replaced by the median of the neighbouring pixels. [6]

3.5 Segmentation

Devanagari characters segmentation process has been depicted as follows. The pre-processed image has been taken as an input image .The most dominating line is called shirorekha [1]. Shirorekha connects all the alphabets together. Remove the shirorekha from the word. Extract the sub images that are separated from the adjacent letters. The sub image may contain more than one component called as modifier. Separate the modifiers if present. Segmented image is sent for feature extraction [3][5][6][7][10].



a)Pre-processed Marathi word AMAN with Shirorekha



b) Pre-processed Marathi word AMAN after removing Shirorekha

c) Character Separation in the Marathi word AMAN

d) Segmentation of Marathi word AMAN

Fig.3**:** Segmentation Process of Marathi word AMAN

3.6 Feature Extraction

Three classes of core characters are decided based on the coverage of core stripe and bar present

End Bar
Middle Bar
No Bar Characters

Twenty one features are extracted and applied to the neural network. End points, Junctions , loops are structural features and aspect ratio is statistical feature. Suppose the Character is अ Then the matrix is

0	0	1
0	0	0
0	0	1

0	0	0
0	0	1
0	0	0

Fig.4:a)End Points b) junction points

Loop is present with Shirorekha and absent without Shirorekha.And the aspect ratio which i simage height/width ratio [1 0 0] So the total code feature vector is

1 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0

If the character is म Then the matrix is

1	0	1
1	0	0
0	0	1

0	0	0
1	0	1
0	0	0

Fig.5: a)End Points b) junction points Loop is present and the aspect ratio is [1 1 0] So the total code feature

1 0 1 1 0 0 0 0 1 0 0 0 1 0 1 0 0 0 1 1 0

If the character is न Then the matrix is

0	1	0
0	0	0
0	0	1

0	0	1
0	1	0
0	0	0

Fig.6: a)End Points b) junction points

Loop is present Aspect ratio is 110

0 1 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 1 1 0

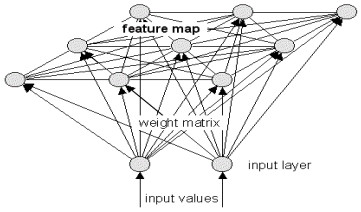
Like this we can calculate the code vector for all Devanagari characters [13]. These features are applied as an input vector to the artificial neural network. The calculated code vector is applied the Classifier i.e. self organizing map.

3.7 Kohonen Feature Map.

Self Organizing map is Topology preserving Map. It has property of human brain which is not observed in many artificial neural network classifiers. It used the concept of closet cluster wins(minimum Euclidean distance) and efficiently clusters the data .Generalization is the unique property of the said classifier.

3.7.1 Basic Structure

The first layer of the network is an input layer. Fig 7 shows connections from input vector to a single unit in the competitive layer Typically the second layer is organized as a two dimensional grid.[4] All the connections go from the first layer to the second layer. Two layers are fully interconnected [13] .



$[x_1,.....x_n]$

Fig.7: Self Organizing Maps. [4]

When an input pattern is presented,[13] each unit in the first layer takes the values of the corresponding entry, from an input pattern. The second layer units then sums the inputs and compute to find a single winning unit. Every unit is associated with its weight value which is initially random. The overall operation of the Kohonen network is similar to the competitive learning paradigm [4]. Each input pattern vector is uniformly distributed between 0 and 1.We can apply input pattern vector with n entries. As a result the input pattern is uniformly spread over d dimensional hypercube. If n = 2, then the input pattern are uniformly spread over a square [4].

The input pattern to the Kohonen pattern is denoted as [4]

$X= [x_1, x_2, x_3, x_4, ---x_n,]$ (1)

The weights are given by

$W= [w_1, w_2, w_3, w_4, -- w_n,]$ (2)

Where W identifies the units in the competitive layer. The first step in the operation of the Kohonen network is to compute a matching value for unit i is [4]

$|| X – W_i ||$ (3)

which is the distance between vector E and Ui

$\sum (x_J – w_{ij})2$ (4)

The unit with the lowest matching value (the best match) wins the competition. Here we denote The unit with the best match as unit c_i and c is chosen such that

$|| X– W_i ||$ = min { $|| X – W_i ||$ } (5)

Where minimum is taken over all the units i in the competitive layer. If two units have same matching value then by convention, the unit with the lower index value i is chosen. After the winning unit is identified, the next step is to identify the neighborhood around it.The neighborhood in this case consists of the units that are within the square that is centered on the winning unit c. The size of the neighborhood changes as shown by squares of different sizes in figure 8. The neighborhood is denoted by set of units N_c. Weights are updated for all neurons that are in the neighborhood of the winning unit .

1 2 3 4 5
6 7 8 9 10
11 12 13 14 15
16 17 18 19 20
21 22 23 24 25

Fig.:8 Network Neighborhood

(By considering 13 th neuron as centre c (winner) we can define the neighborhood as

N_1 = { 7 8 9 12 14 17 18 19}

N_2 = { 2 3 4 6 10 11 15 16 20 22 23 24 })

The update equation is

$$Wij = \{ \sum (x_j - w_{ij}) \} \quad (6)$$

If the unit i is the neighborhood Nc otherwise And

$$w_{ijnew} = w_{ij}^{old} + w_{ij} \quad (7)$$

The adjustments results in the winning unit and its neighbors having their weights modified becomes more like the input pattern[13]. The winner then becomes more likely to win the competition. Two parameters that should be specified, the value of α_i, learning rate parameter and neighborhood size Nc, the learning rate, [4]

is specified as

$$\alpha = \alpha_0 \{1 - (t/T)\} \quad (8)$$

Where t = the current training iteration

T = Total no. of training iterations to be done .

Thus α_0 is decreased until it reaches 0. the decrease is linear with number of training iterations completed The size of neighborhood is second parameter to be specified. Typically the initial neighborhood width is relatively large and the width is decreased over many training iterations. Consider the neighborhood as shown in fig 8 which is centered on winning unit c, at position (x_c, y_c). Let d be the distance from c to the edge of the neighborhood. The neighborhood is then all (x,y) Such that

$$C-D < X < C+D \text{ And } C-D < Y < C+D \quad (9)$$

This defines a square neighborhood about c. Since the width of the neighborhood decreases over the training iterations, the value of d decreases. Initially d is set at a chosen value denoted by d_0 may be chosen at a half or a third of the width of the competitive layer of the processing units[][4]. The value of d is then made to decrease according to the equation [4] $d = d_0 \{1 - (t/T)\}$ (10)

Where t = the current training iteration

T = Total no. Of training iterations to be done. This process assures a gradual linear decrease in d, starting with d_0 and going down to 1. The same amount of time is spend at each value.

4. PERFORMANCE ANALYSIS

The scheme is implemented and discussed as below. The database is prepared from 500 different persons. Each image is scanned and segmented . Feature vector is applied to the classifier.
Accuracy has been calculated by using the formula.

$$\text{Accuracy} = \frac{T_P \text{ (True Positive)} + T_N \text{ (True Negative)}}{T_P + T_N + F_P \text{ (False Positive)} + F_N \text{ (False Negative)}}$$

Table 1 Percentage Accuracy Table

Character	No. of Samples for Training	No. of Samples for Testing	Accuracy With SOM Classifier.
अ	**1500**	1000	99
आ	**1500**	1000	99
इ	**1500**	1000	99
ई	**1500**	1000	88
उ	**1500**	1000	99
ऊ	**1500**	1000	85
ऋ	**1500**	1000	93
ए	**1500**	1000	99
ऐ	**1500**	1000	87
ओ	**1500**	1000	87
औ	**1500**	1000	83
अं	**1500**	1000	83
अः	**1500**	1000	82
क	**1500**	1000	99
ख	**1500**	1000	94
ग	**1500**	1000	90
घ	**1500**	1000	99
ङ	**1500**	1000	90
च	**1500**	1000	99
छ	**1500**	1000	97
ज	**1500**	1000	99
झ	**1500**	1000	99
ञ	**1500**	1000	99
ठ	**1500**	1000	99
ड	**1500**	1000	99
ढ	**1500**	1000	99
ण	**1500**	1000	99
त	**1500**	1000	99
थ	**1500**	1000	99
द	**1500**	1000	99
ध	**1500**	1000	99
न	**1500**	1000	99
प	**1500**	1000	99
फ	**1500**	1000	99
ब	**1500**	1000	99
भ	**1500**	1000	99
म	**1500**	1000	99
य	**1500**	1000	99
र	**1500**	1000	99

ल	**1500**	1000	99
व	**1500**	1000	99
श	**1500**	1000	80
ष	**1500**	1000	99
स	**1500**	1000	99
ह	**1500**	1000	99
ळ	**1500**	1000	99
क्ष	**1500**	1000	99
ज्ञ	**1500**	1000	99
श्र	**1500**	1000	99

Table 2 Percentage Accuracy Table

Numeral	No. of Samples for Training	No. of Samples for Testing	Accuracy With SOM Classifier.
०	**1500**	1000	99
१	**1500**	1000	99
२	**1500**	1000	99
३	**1500**	1000	99
४	**1500**	1000	99
५	**1500**	1000	99
६	**1500**	1000	99
७	**1500**	1000	99
८	**1500**	1000	99
९	**1500**	1000	99

Table 3 Final Classification of Marathi Characters

Classification of the characters	Example
Open header letter with End Bar	अ, थ, ध ,भ
One conjunction letter with End Bar	च ,ज ,त ,ञ ,न , ज्ञ ,व
More conjunction letter with End Bar	ख ,घ ,झ ,प ,म ,य ,ष , स
Middle Bar	क ,फ ,ऋ
No Bar	इ ,ठ ,र ,ळ ,ड ,ढ ,द , ए
Special Case	ग ,श , ण

5. CONCLUSION AND DISCUSSIONS

The self-organizing map is used for pattern classification. The performance achieved by self- organizing map is better than back propagation algorithm. Generally back propagation algorithm is used for such tasks. The disadvantages associated with back propagation algorithm such as local minima and deciding number of hidden units is not observed in the implemented system. Self organizing map is trained for different parameters the classification accuracy obtained is 85% to 95% for 300 nodes, except some special *characters*. The training time is proportional to the number of patterns used for training, number of output nodes and iterations. After fine tuning the accuracy is increased by 2% . The data base from 2000 persons is taken for training and for testing 500 samples is taken per character from different persons. In SOM initially one hidden layers taken as the training progresses, observed .winner index of different classes are either same or closer to the previous winner index. The characters are divided into three different classes based on whether the character is no bar, Middle bar, end bar . The output is fine tuned for different classes.

6. ACKNOWLEDGMENTS

The authors are thankful to Prof. Dr. A.R.Bhalerao ,Dean, Engineering and Technology, Bharati Vidyapeeth Deemed University,Pune for providing the infrastructure , continuous inspiration and constant support.

7. REFERENCES

[1] A..M.Namboodiri, A.K.Jain, "Online script recognition," IEEE Trans. Patten Analysis and Intelligence, Vol.26, No.1, pp.124-130,January 2004

[2] U.Bhattacharya,B.B.Choudhari, "Handwritten Numeral database of Indian Scripts and Multistage Recognition of Mixed Numbers" IEEE Trans. Patten Analysis and Machine Intelligence, Vol.31No.3 pp. 444-457, March.2009.

[3] In-Jung Kim and Jin Hyung Kim,"Statistical character structure modeling and its application to handwritten Chinese recognition," Pattern Analysis and Machine Intelligence, Vol.25, No.11, pp.1422-1436, 2003.

[4] T.Kohonen, "The Self Organizing Map," IEEE Trans. Patten Analysis and Intelligence, Vol.78, No.9 pp. 1464-1480, Sept.1990.

[5] Macro Bressan and Jordi Vitria , " On the selection and Classification of Independent Features," IEEE Trans. Patten Analysis and Intelligence, Vol.25 No.10pp. 1312-1322, October 2003.

[6] A.W. Senior, J.Robinson,"An Offline Cursive Handwriting Recognition System,"IEEE Trans. Patten Analysis and Intelligence, Vol.20 No.3 pp. 309-321, March 1998.

[7] B.Wegmann, C.Zetzshe, "Feature – Specific Vector Quantization Of Images," IEEE Trans.Image Processing, Vol.5 No.2 pp. 274-288, Feb. 2000.

[8] J.A.Starzyk, Zhen Zhu, "Self – Organizing Learning Array," IEEE Trans Neural Network, Vol.16 No.2 pp. 355-363 March 2005.

[9] J.Park,"An Adaptive Approach to Offline Handwritten Word Recognition,"IEEE Trans. Patten Analysis and Intelligence, Vol.24 No.7 pp. 919-931, July.2002.

[10] Cheng- Lin Liu , S. Jaeger,M. Nakagawa, "Online Recognition of Chinese Characters: The State – Of – the Art," IEEE Trans. Patten Analysis and Intelligence, Vol.26 No.2pp. 198-213,February 2004.

[11] Luiz S. Oliveria,F. Bortolozzi,R.Sabourn "Automatic Recognition of Handwritten Numeral Strings:A Recognition and Verification Strategy ," IEEE Trans. Patten Analysis and Intelligence, Vol.24 No.11 pp. 1438-1453, NOV.2002.

[12] Reena Bajaj, Lipika Day, Santanu Chaudhari, "Devanagari Numeral Recognition by Combining Decision of Multiple Connectionist Classifiers", Sadhana, Vol.27, Part-I, 59-72, 2002.

[13] V.R.Pawar;GaikwadA.N,"Multistage Recognition Approach for HandwrittenDevanagariScriptRecognition"Topic(s): Communication, Networking & Broadcasting ; Components, Circuits, Devices & Systems ; Computing & Processing(Hardware/Software) DigitalObjectIdentifier:10.1109/WICT.2012.6409156 Publication Year: 2012 , Page(s): 651 - 656 IEEE Conference Publications.