Writer Identification using GMM Supervectors

Nikhil Bansal (20161065)
Mudit Surana (20161100)
Rohit Kumar Agarwal (20161011)

INTRODUCTION

- The paper proposes a new system for writer identification using GMM supervectors
- Writer Identification attempts to identify author of a document, given a known set of authors
- The proposed method uses GMM supervectors to encode the feature distribution of individual writers
- Here, we compare the GMM supervectors approach to other encoding schemes like Fisher vectors and VLAD (Vectors of Locally Aggregated Descriptors)
- Here we have used ICDAR-13 dataset for comparison

GAUSSIAN MIXTURE MODEL

- A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.
- The parameters of GMM are estimated using the EM algorithm
- We used sklearn.mixture package which enables one to learn Gaussian Mixture Models with flag ‘diagonal’.

UNIVERSAL BACKGROUND MODEL

- The UBM is the background model that is used generally in speaker recognition which is created by estimating a GMM from large spectral feature vectors.
- In this paper, the UBM is also created by estimating a GMM but here the model is estimated using a set of SIFT descriptors computed from training documents.
- The paper employed **RootSIFT** features i.e a variant of SIFT where the features are normalized using the square root (Hellinger) kernel.

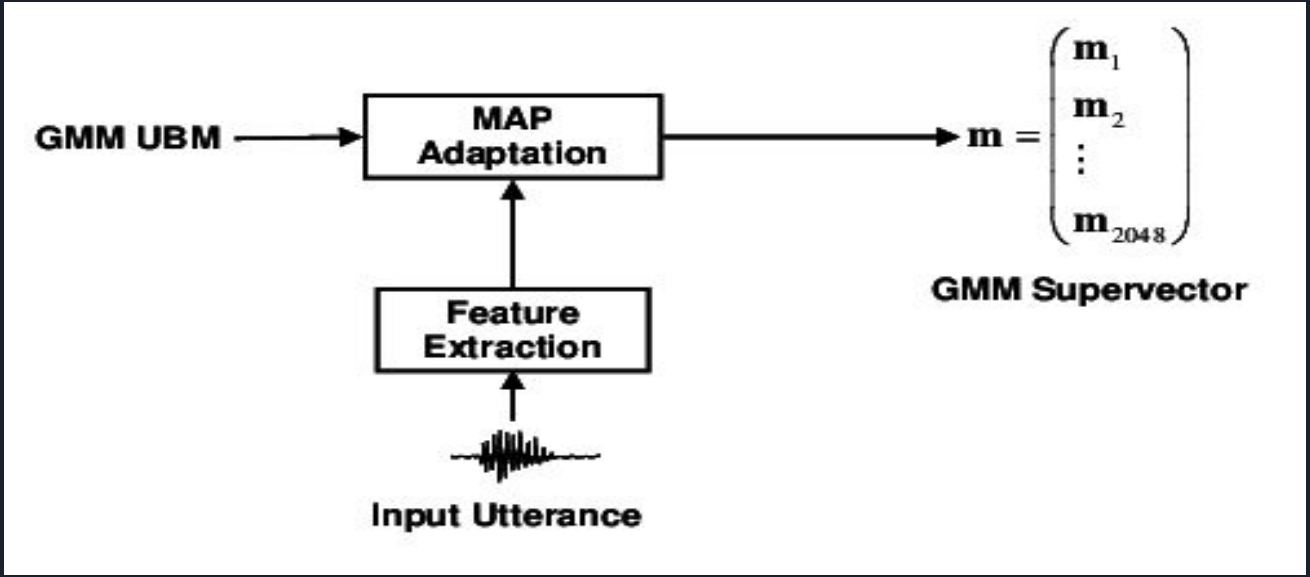
GMM Adaptation

- The final UBM is adapted to each document individually, using all the SIFT descriptors
- For the MAP estimation, first the posteriors probabilities are computed for all mixtures, then the mixture parameters are adapted.
- Mixture with high posteriors are adapted more strongly.
- This is controlled by a fixed relevance factor for the adaptation coefficients.

METHOD

- RootSIFT features for each document are computed
- Descriptors from an independent document are used to train a vocabulary i.e. our UBM (Universal Background Model)
- UBM is adapted to test each document individually
- These new features are stacked into a **supervector** , to form a feature vector for each document

Constructing a GMM supervector


GMM UBM
MAP Adaptation
Feature Extraction
Input Utterance
$\mathbf{m} = \begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \vdots \\ \mathbf{m}_{2048} \end{pmatrix}$
GMM Supervector

OTHER ENCODINGS

FISHER ENCODING

- The Fisher encoding uses GMM to construct a visual word dictionary.

VLAD ENCODING

- The Vectors of Locally Aggregated Descriptors is similar to Fisher vectors but it does not store second-order information about the features.

APPROACH

- RootSIFT features are a variant of SIFT where the features are normalized using the square root (Hellinger) Kernel
- UBM is created by estimating a GMM from a set of SIFT descriptors
- The GMM parameters are estimated using the expectation-maximization algorithm (EM)
- For Normalization, first the square root is computed element-wise and then each vector is L2-normalized

EVALUATION METRICS

TOP-1

- First the distance from query supervector to all other supervectors is computed.
- Resulting distances are then sorted and author with least distance is assigned to the query document

TOP-2

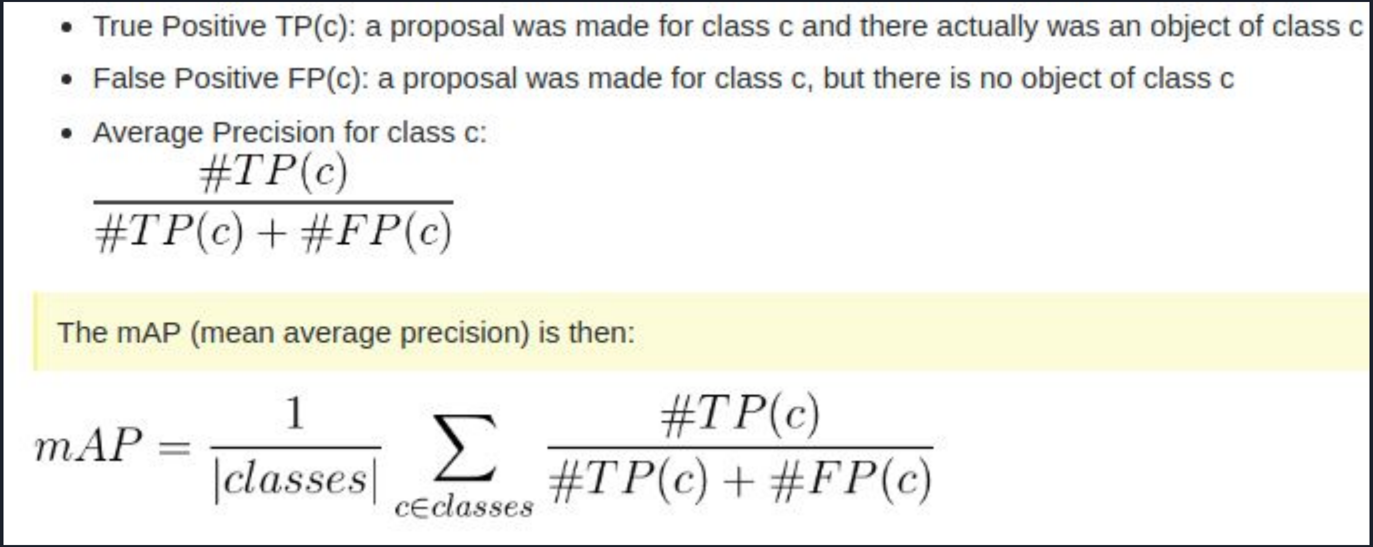
- Same as TOP-1 except that for correct classification, both the top-2 least distant documents (to the query document), should have the same label as query document.

TOP-3

- Same as TOP-1 except that for correct classification, both the top-3 least distant documents (to the query document), should have the same label as query document.

EVALUATION METRICS (continued)

mAP (Mean Average precision)



- True Positive TP(c): a proposal was made for class c and there actually was an object of class c
- False Positive FP(c): a proposal was made for class c, but there is no object of class c
- Average Precision for class c:

$$\frac{\#TP(c)}{\#TP(c) + \#FP(c)}$$

The mAP (mean average precision) is then:

$$mAP = \frac{1}{|classes|} \sum_{c \in classes} \frac{\#TP(c)}{\#TP(c) + \#FP(c)}$$

RESULTS

Supervector encoding outperforms the other two tested encodings, namely Fisher vectors and VLAD

GMM Supervector	0.9725
Fisher	0.955
VLAD	0.915

Table 1. Encodings and their TOP-1 Accuracy

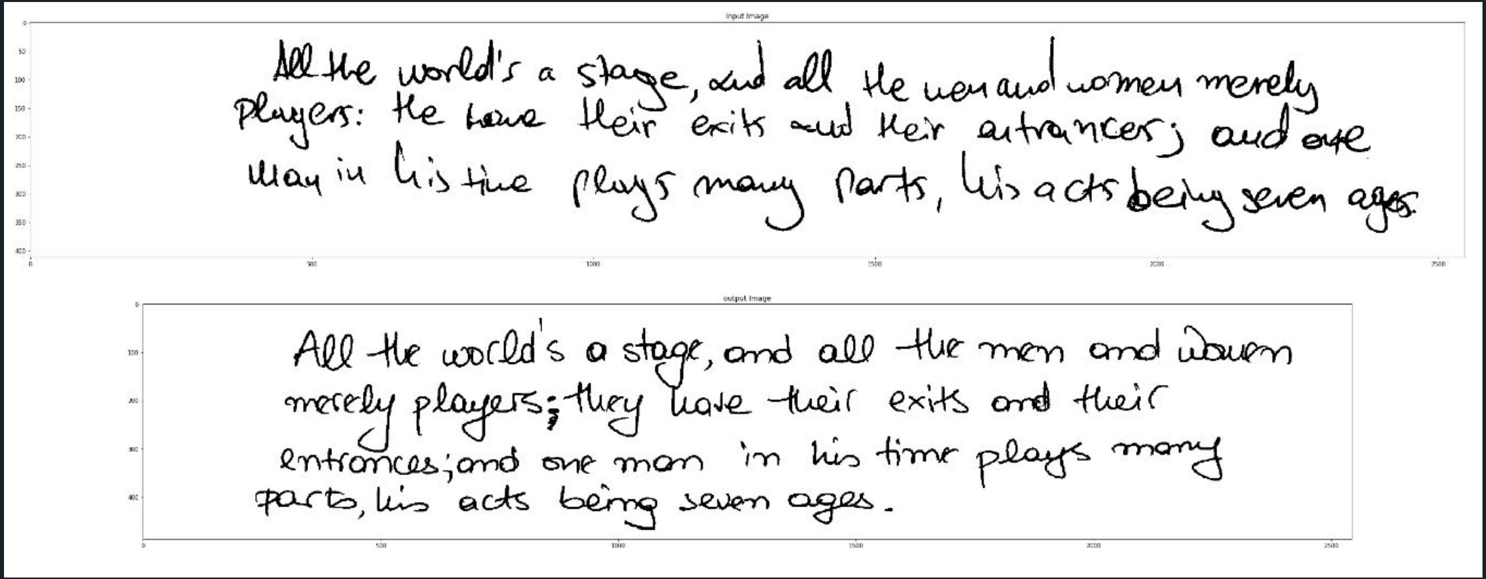
RESULTS (continued)

	TOP-1	TOP-2	TOP-3	mAP
VLAD	0.925	0.68	0.4025	0.7125
FV	0.94	0.72	0.475	0.748
SV	0.9725	0.7575	0.41	0.7525

Table 2. Comparison between different encodings using different evaluation metrics

RESULTS (continued)

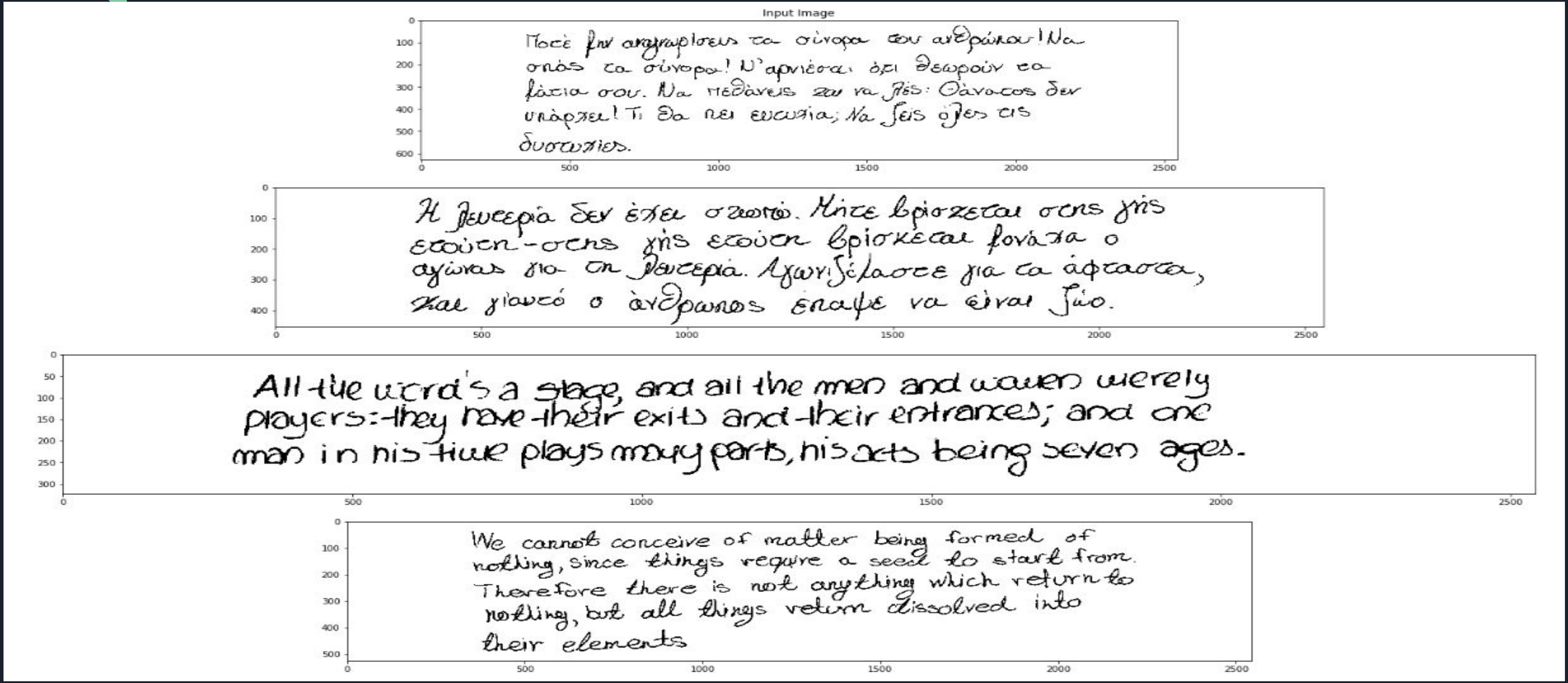
Incorrect classification on a document in ICDAR 13, using GMM supervector model


Input Image
output Image

All the world's a stage, and all the men and women merely
Players: the have their exits and their entrances; and one
man in his time plays many parts, his acts being seven ages.

All the world's a stage, and all the men and women
merely players; they have their exits and their
entrances; and one man in his time plays many
parts, his acts being seven ages.

RESULTS: The below image shows the documents from ICDAR 13 dataset, the first one is the input image and the other three are the top-3 outputs(as per the cos-distance between documents), the second one has the incorrect label:


Input Image

Ποτέ μην αναγνωρίσεις τα σύνορα του ανθρώπου! Να σπάς τα σύνορα! Ν' αρνιέσαι ό,τι θεωρούν τα μάτια σου. Να πεθάνεις και να ζήσεις: Θάνατος δεν υπάρχει! Τι θα πει ευτυχία; Να ζεις όλες τις δυστυχίες.

Η λευτεριά δεν έχει σκοπό. Μήτε βρίσκεται στης γης στοιχειό - στης γης στοιχειό βρίσκεται μονάχα ο αγώνας για τη λευτεριά. Αγωνιζόμαστε για τα άφταστα, και γι' αυτό ο άνθρωπος έπαψε να είναι ζώο.

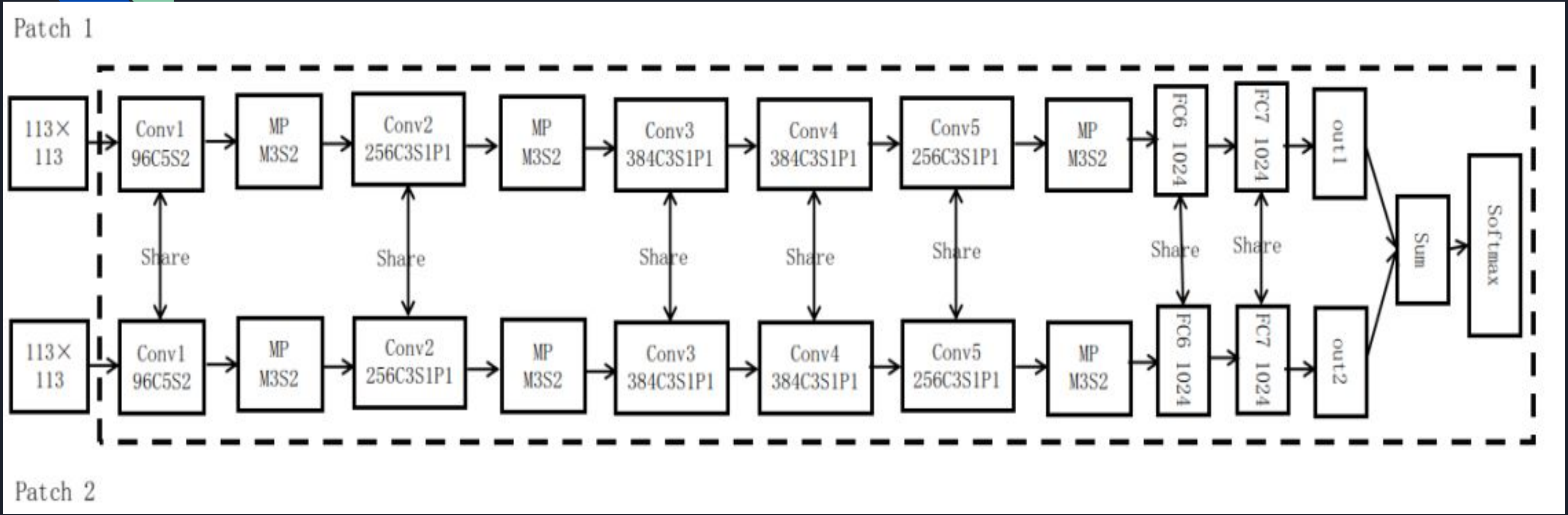
All the word's a stage, and all the men and women merely players:-they have their exits and their entrances; and one man in his time plays many parts, his acts being seven ages.

We cannot conceive of matter being formed of nothing, since things require a seed to start from. Therefore there is not anything which return to nothing, but all things return dissolved into their elements

WRITER IDENTIFICATION USING DEEP CNN

- We design and optimize multi-stream structure for Writer Identification task
- We use data augmentation learning to enhance the performance of the Deep CNN
- We use a patch scanning strategy to handle text image with different lengths
- The network takes multiple local regions as input and is trained with softmax loss on identification
- Here, we use IAM dataset for training and evaluation

NETWORK ARCHITECTURE



INPUT

- The input to the model are not unique sentences but rather random patches cropped from each sentence.
- Resize each sentence so that new height is 113 pixels and new width is such that original aspect ratio is maintained, since, distorting the shape of image by changing the aspect ratio resulted in a big drop in model performance.
- From the adjusted image, patches of 113x113 are randomly cropped, which are then given as input to the model.

TESTING

- Scan the testing image to generate image patches
- Compute the final score for each writer by averaging scores of all image patches for that writer
- Return the writer with the highest score

RESULTS

Following table shows the accuracy comparison between the CNN results and the other 3 encodings namely Supervector, Fisher and VLAD.

CNN	0.988
GMM Supervector	0.9725
Fisher	0.94
VLAD	0.925

Table 3. Comparison between different models