Word Spotter

Ritvik Agrawal
Vivek Chandela
Kirandevraj R

Objective

Find all instances of a given word in potentially a large dataset of document images in a multi-writer setting.

Queries:

1. Query by example (Image)
2. Query by string (Text)

Challenges with previous works:

Most popular techniques are based on describing word images as sequence of features of variable length and uses Dynamic Time Warping(DTW) technique to classify them.

- Out Of Vocabulary(OOV) spotting (words not in training data but present in testing data)
- Time taken for image retrieval
- Same word, different handwritings

Current Approach

- Instead of learning models for keywords, learning what makes words and letters unique independently of their writers style
- Using attribute based representation for each word

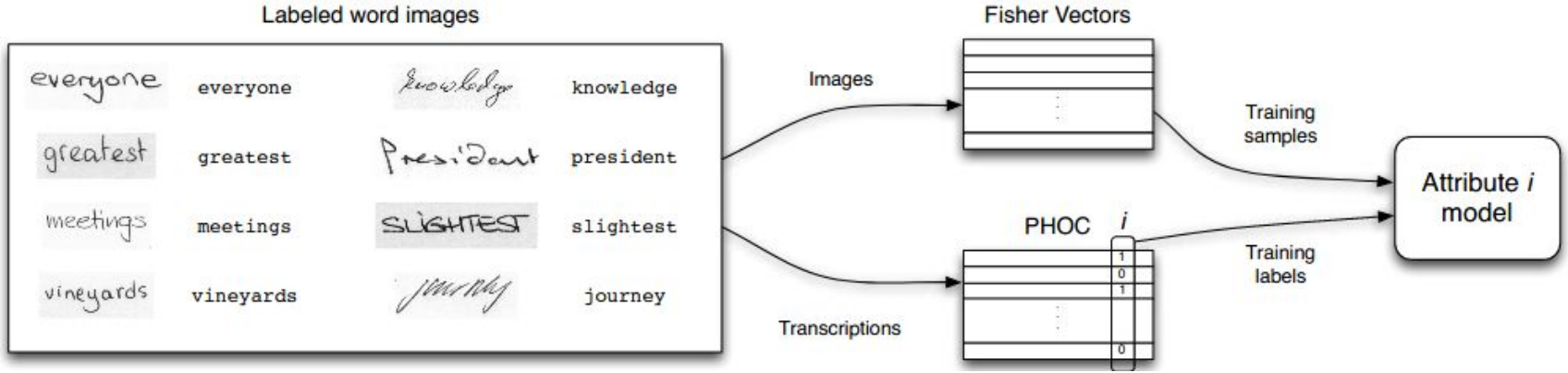


Image Representation

We use fisher vector as our base representation of images.

- Image is segmented into 2 x 6 segments
- SIFT features are densely extracted from the images over a 2x6 spatial grid and reduced to 64 dimensions with PCA
- 16 GMMs are trained on these images separately for each segment positions.
- These SIFT features are aggregated into a FV(Fisher Vector) which considers the gradients with respects of the means and variances of the trained GMM model.
- Final representation size: 2 x 16 x 64 x 12 = 24,576 dimension for a single image

String Representation

We use PHOC(pyramidal histogram of characters) to represent strings

- The binary histogram encodes whether a character is present or not
- We use levels 2, 3 and 4 as well as 75 common bigrams at level 2, leading to 384 dimensions
- Spatial pyramid representation ensures that the information on characters order is preserved

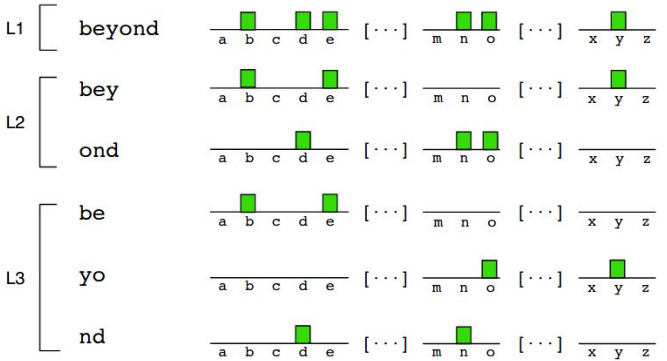
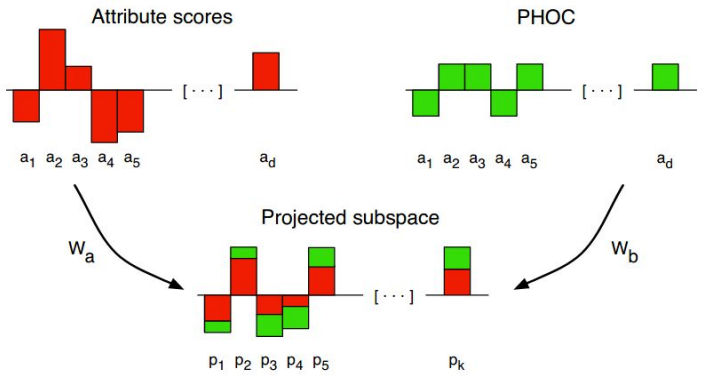


Figure 1. PHOC histogram at levels 1, 2, and 3. The final PHOC histogram is the concatenation of these partial histograms.

Projecting into common subspace

We use canonical correlation analysis (CCA) scheme to project the image representation and string representation into a common subspace.

- The goal of CCA is to find a projection of each view that maximizes the correlation between the projected representations.
- The projected subspace has a representation of dimension 192



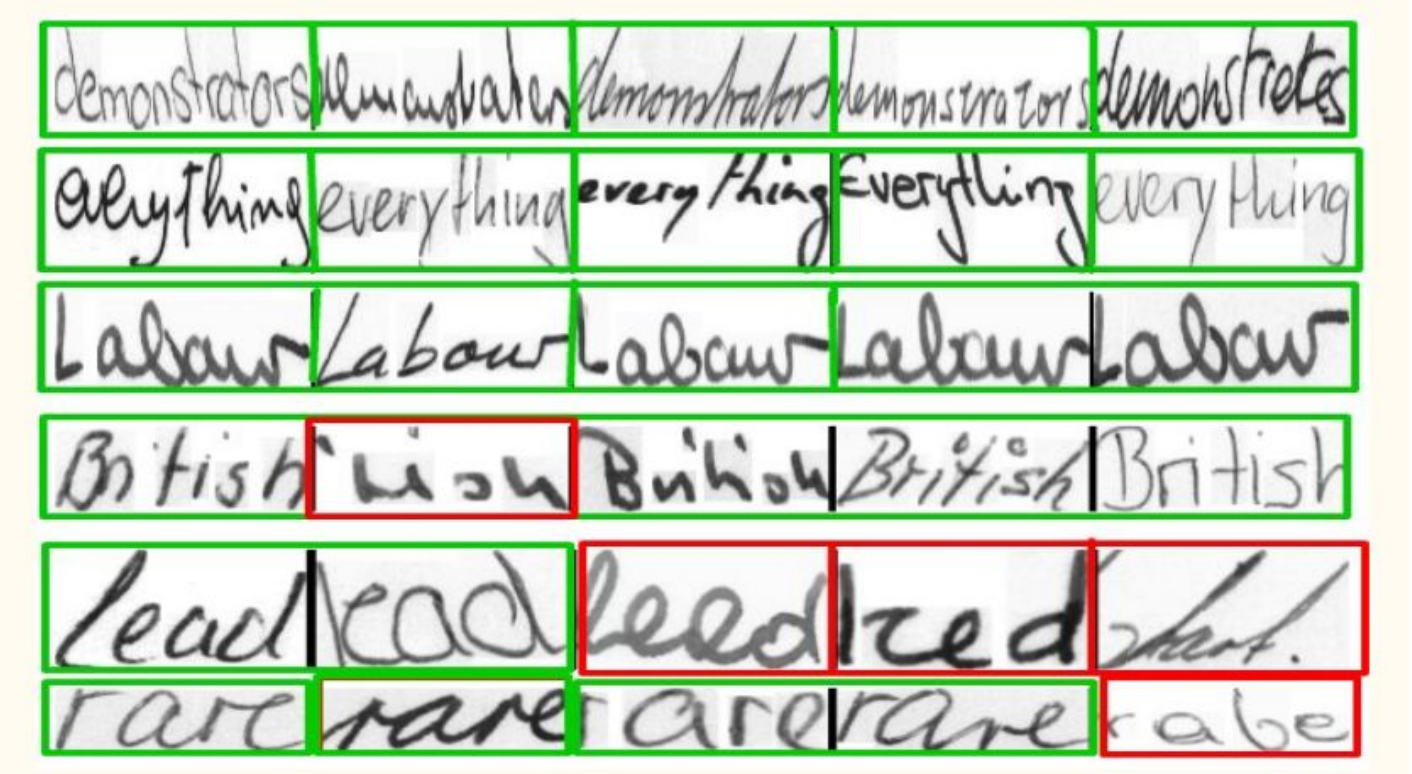
Results

- Two representations in the projected subspace are compared by cosine similarity
- The performance is measured in terms of mean average precision.

	FV	FV + CCA
QBE	14.00	48.00
QBS	-	37.00

Sample output - QBE

Sample output - QBS



Implementation Challenges and Observations

- The IAM dataset contains 115,000 word images with transcriptions and hence required huge computing resources during implementation
- The dataset was divided into 40 / 40 / 20 percent for training the PCA, CCA and testing
- Tried different implementations of GMM and FV calculations and used optimized one.
- The performance of the word spotter increases as we increase the dimension size of the projected subspace from 192 to 256

Work done

- Images to FV implementation [Ritvik, Vivek, Kiran]
- PHOC implementation [Kiran]
- CCA implementation [Vivek]
- MAP implementation [Ritvik]

Github Link:
https://github.com/Kirandevraj/Handwritten-Word-Spotting-with-Corrected-Attributes

Thank you