A Hierarchical Codebook Descriptor Approach for Online Writer Identification

Vivek Venugopal, Surbhi Pillai, Suresh Sundaram
EEE department, Indian Institute of Technology, Guwahati

Contribution

- Derive a codebook based descriptor which reduces the dimension while providing comparable results to [1, 2] .
- Proposed descriptor has a dimension independent of the size of the feature vector.

Codebook Descriptor

Proposed codebook descriptor

Given a hierarchical codebook $\{\{\mathbf{c}_{ij}\}_{j=1}^{K_2}\}_{i=1}^{K_1}$ and feature vectors $\{\mathbf{f}^j\}_{j=1}^{N_T}$ from a document having N_T points, each feature vector is assigned to the nearest codevector based on the minimum Euclidean distance criterion. Let $\{\mathbf{f}_{ij}^p\}_{p=1}^{n_{ij}}$ denote the feature vectors assigned to codevector $\mathbf{c}_{ij}$ where $n_{ij} < N_T$ and $\sum_{i=1}^{K1}\sum_{j=1}^{K2} n_{ij} = N_T$.

$$S_{ij}^{p+}(d) = \begin{cases} \frac{1}{1+|f_{ij}^p(d)-c_{ij}(d)|} & f_{ij}^p(d) \geq c_{ij}(d) \\ 0 & otherwise \end{cases}$$

$$S_{ij}^{p-}(d) = \begin{cases} \frac{-1}{1+|f_{ij}^p(d)-c_{ij}(d)|} & f_{ij}^p(d) < c_{ij}(d) \\ 0 & otherwise \end{cases}$$

$$1 \leq p \leq n_{ij}, 1 \leq d \leq D$$

$$\tilde{S}_{ij}^{+}(d) = \frac{\sum_{p=1}^{n_{ij}} S_{ij}^{p+}(d)}{\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{p=1}^{n_{ij}} S_{ij}^{p+}(d)}$$

$$\tilde{S}_{ij}^{-}(d) = \frac{\sum_{p=1}^{n_{ij}} S_{ij}^{p-}(d)}{\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{p=1}^{n_{ij}} S_{ij}^{p-}(d)}$$

$$1 \leq d \leq D, 1 \leq i \leq K_1, 1 \leq j \leq K_2$$

$$\tilde{\mathbf{S}}_{ij}^{+} = [\tilde{S}_{ij}^{+}(1)....\tilde{S}_{ij}^{+}(d)....\tilde{S}_{ij}^{+}(D)]$$

$$\tilde{\mathbf{S}}_{ij}^{-} = [\tilde{S}_{ij}^{-}(1)....\tilde{S}_{ij}^{-}(d)....\tilde{S}_{ij}^{-}(D)]$$

$$\tilde{\mathbf{S}}_{ij} = [\| \tilde{\mathbf{S}}_{ij}^{+} \|_2 \| \tilde{\mathbf{S}}_{ij}^{-} \|_2]^T$$

Proposed Methodology

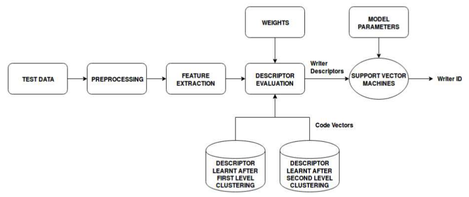

Figure 1: Block diagram of proposed online writer identification system.

Weights formulation

Weights formulation

1. Generation of histogram for each subcluster $\mathbf{c}_{ij}$.
2. Entropy computation for the generated histogram in (1).
3. Calculation of weights as a function of computed entropy.

$\mathbb{H}_{ij}$ be the histogram for the codevector $\mathbf{c}_{ij}$ with R bins $\{v_{ij}^1, v_{ij}^2, ...v_{ij}^R\}$ where R is number of writers considered for codebook generation.

$$\tilde{h}_{ij}^r = \frac{v_{ij}^r}{n_{ij}^r}$$

$$h_{ij}^r = \frac{\tilde{h}_{ij}^r}{\sum_{r=1}^{R} \tilde{h}_{ij}^r}$$

$$H_{ij} = \sum_{r=1}^{R} -h_{ij}^r \log_2 h_{ij}^r$$

$$w_{ij} = \frac{1}{1+H_{ij}}$$

Proposed writer descriptor

$$\mathbf{S}_{ij} = w_{ij} \times \tilde{\mathbf{S}}_{ij}$$

$$\mathbf{S} = [\mathbf{S}_{11} \quad \mathbf{S}_{12} \quad ... \quad \mathbf{S}_{ij} \quad ... \quad \mathbf{S}_{K_1K_2}]^T$$

Feature Extraction

- **Speed**(1):
- **Writing Direction**(2): The cosine and sine of the angle θ_i that the vector $\mathbf{p}_i - \mathbf{p}_{i+r}$ makes with the horizontal.
- **Curvature**(2): cosine and the sine of the angle ϕ_i, defined between the vectors $\mathbf{p}_{i+r} - \mathbf{p}_i$ and $\mathbf{p}_i - \mathbf{p}_{i-r}$ respectively.
- **Vicinity aspect**(1) : height to width ratio of the bounding box BB encompassing the points $\{\mathbf{p}_{i-r}, ..., \mathbf{p}_i, ..., \mathbf{p}_{i+r}\}$.
- **Vicinity Curliness**(1): is the ratio of the trajectory length to the maximum amongst width and height of BB.

Results

Table 1: Average identification rates with varying values of gap parameter.

	Paragraph level		Textline level	
r	IR	(K_1, K_2)	IR	(K_1, K_2)
1	98.61	(4, 40)	81.96	(5, 50)
2	98.77	(4, 50)	86.03	(5, 55)
3	**98.85**	(4, 50)	**89.62**	(5, 60)
4	98.18	(4, 40)	88.37	(5, 60)
5	98.02	(4, 40)	86.01	(5, 50)
6	97.77	(4, 50)	84.89	(5, 70)

Table 2: Survey of online writer identification system on IAM database.

Methodology	Paragraph Level	Textline Level
GMM based system [3]	98.56	88.96
Latent Dirichlet Allocation [4]	93.39	-
Subtractive Clustering + tf-idf scoring [5]	96.30	-
Sparse + tf-idf scoring [6]	98.94	83.3
K-means + Codebook descriptor [1]	97.81	80.61
Improved codebook descriptor [2]	98.82	**89.92**
Proposed descriptor	**98.85**	89.62

References

[1] Vivek Venugopal and Suresh Sundaram. An online writer identification system using regression-based feature normalization and codebook descriptors. *Expert Systems with Applications*, 72:196 – 206, 2017.

[2] "Vivek Venugopal and Suresh Sundaram". An improved online writer identification framework using codebook descriptors. *Pattern Recognition*, 78:318 – 330, 2018.

[3] Andreas Schlapbach, Marcus Liwicki, and Horst Bunke. A writer identification system for on-line whiteboard data. *Pattern Recogn.*, 41(7):2381–2397, July 2008.

[4] A. Shivram, C. Ramaiah, and V. Govindaraju. A hierarchical bayesian approach to online writer identification. *IET Biometrics*, 2(4):191–198, December 2013.

[5] G. Singh and S. Sundaram. A subtractive clustering scheme for text-independent online writer identification. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 311–315, Aug 2015.

[6] I. Dwivedi, S. Gupta, V. Venugopal, and S. Sundaram. Online writer identification using sparse coding and histogram based descriptors. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 572–577, Oct 2016.