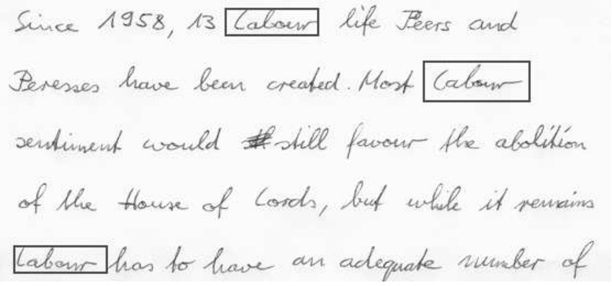
Script Independent Word Spotting in Multilingual Documents

Anurag Bhardwaj, Damien Jose and Venu Govindaraju
Center for Unified Biometrics and Sensors (CUBS)
University at Buffalo, State University of New York
Amherst, New York 14228
{ab94,dsjose,govind}@cedar.buffalo.edu

Abstract

This paper describes a method for script independent word spotting in multilingual handwritten and machine printed documents. The system accepts a query in the form of text from the user and returns a ranked list of word images from document image corpus based on similarity with the query word. The system is divided into two main components. The first component known as Indexer, performs indexing of all word images present in the document image corpus. This is achieved by extracting Moment Based features from word images and storing them as index. A template is generated for keyword spotting which stores the mapping of a keyword string to its corresponding word image which is used for generating query feature vector. The second component, Similarity Matcher, returns a ranked list of word images which are most similar to the query based on a cosine similarity metric. A manual Relevance feedback is applied based on Rocchio's formula, which re-formulates the query vector to return an improved ranked listing of word images. The performance of the system is seen to be superior on printed text than on handwritten text. Experiments are reported on documents of three different languages: English, Hindi and Sanskrit. For handwritten English, an average precision of 67% *was obtained for* 30 *query words. For machine printed Hindi, an average precision of* 71% *was obtained for* 75 *query words and for Sanskrit, an average precision of* 87% *with* 100 *queries was obtained.*



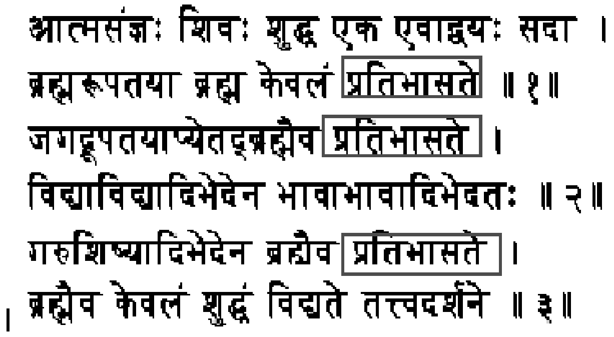
Since 1958, 13 Labour life Peers and
Peresses have been created. Most Labour
sentiment would ~~#~~ still favour the abolition
of the House of Lords, but while it remains
Labour has to have an adequate number of

Figure 1: A Sample English Document - Spotted Query word shown in the bounding box.

1 Introduction

The vast amount of information available in the form of handwritten and printed text in different languages poses a great challenge to the task of effective information extraction. Research in this area has primarily focussed on OCR based solutions which are adequate for Roman Language (A sample English document is shown in Figure 1). However efficient solutions do not exist for scripts like Devanagari. One of the main reasons for this is lack of generalisation. OCR solutions tend to be specific to script type. Ongoing research continues to scale these methods to different types and font sizes. Furthermore, non-Latin scripts exhibit complex character classes (like in the Sanskrit document shown in Figure 2) and poor quality documents are common.

The notion of Word spotting [6] has been introduced as an alternative to OCR based solutions. It can be defined as an information retrieval task that finds all occurences of a typed query word in a set of handwritten



आत्मसंज्ञः शिवः शुद्ध एक एवाद्वयः सदा ।
ब्रह्मरूपतया ब्रह्म केवलं प्रतिभासते ॥ १ ॥
जगद्रूपतयाप्येतद्ब्रह्मैव प्रतिभासते ।
विद्याविद्यादिभेदेन भावाभावादिभेदतः ॥ २ ॥
गरुशिष्यादिभेदेन ब्रह्मैव प्रतिभासते ।
। ब्रह्मैव केवलं शुद्धं विद्यते तत्त्वदर्शने ॥ ३ ॥

Figure 2: A Sample Sanskrit Document - Spotted Query word shown in the bounding box.

or machine printed documents. While spotting words in English has been explored [3, 5, 4, 7, 11], generalising these approaches to multiple scripts is still an ongoing research task. Harish et.al [1] describe a 'Gradient, Structural, Concavity' (GSC) based feature set for word spotting in multiple scripts. However, they do not report the average precision rate for all queries in their experimental results which makes it difficult to estimate the performance of their methodology.

One important factor in finding a script independent solution to word spotting is use of image based features which are invariant to script type, image scale and translations. This paper proposes the use of moment based features for spotting word images in different scripts. We describe a moment-function based feature extraction scheme and use the standard vector space model to represent the word images. Similarity between the query feature vector and the indexed feature set is computed using a cosine similarity metric. We also apply the Rocchio formula based Manual Relevance feedback to improve the ranking of results obtained. We evaluate the performance of our system by conducting experiments on document images of three different scripts: English, Hindi and Sanskrit.

The organization of the rest of the paper is as follows: Section 2 describes the previous work. Section 3 describes the theory of moment functions. Section 4 describes indexing word images and feature extraction. Section 5 describes the Similarity Matching and Relevance Feedback method applied to re-rank results. Section 6 describes the experiments and results. Future work and conclusions are outlined in Section 7.

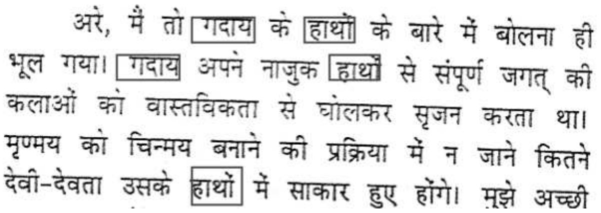
2 Previous Work

Spotting words in English has recently received considerable attention. Manmatha et al. [7], have proposed a combination of feature sets well suited for this application. For finding similarity between a query word image and the document word image, Dynamic Time warping [8] is commonly used. Although the approach has been promising with English handwritten documents, it does not generalise well across scripts. For eg., presence of Shirorekha in Devanagari script (an example shown in Figure 3) renders most of the profile based features ineffective. Also, DTW based approaches are slow. Approaches which use a filter based feature set [2], are efficient with uniform font size and type but are not able to handle font variations and translations.

Harish et al. [1] use a Gradient, Structural and Concavity (GSC) feature set which measures the image characteristics at local, intermediate and large scales. Features are extracted using a $4x8$ sampling window to gather information locally. Since character segmentation points are not perfectly located, local information about stroke orientation and image gradient is not sufficient to characterize the change in font scale and type. Moreover, presence of noise in small regions of the word image lead to inconsistency in the overall feature extraction process. The performance of their approach is presented in terms of percentage of the number of times the correct match was returned, which does not capture the recall rate of system. For English word spotting, their results do not state the size of the dataset and precision recall values have been reported for only 4 query words. For Sanskrit word spotting, the total number of query words is not mentioned which makes understanding of precision recall curve difficult. A comparison of their results against our proposed method is presented in section 6.

3 Moment Functions

Moments and functions of moments have been previously used to achieve an invariant representation of a two-dimensional image pattern [9]. Geometrical moments



अरे, मैं तो गदाय के हाथों के बारे में बोलना ही
भूल गया। गदाय अपने नाजुक हाथों से संपूर्ण जगत् की
कलाओं को वास्तविकता से घोलकर सृजन करता था।
मृण्मय को चिन्मय बनाने की प्रक्रिया में न जाने कितने
देवी-देवता उसके हाथों में साकार हुए होंगे। मुझे अच्छी

Figure 3: A Sample Hindi Document - Spotted Query words shown in the bounding box.

[9] have the desirable property of being invariant under the image translation, scale and stretching and squeezing in either X or Y direction. Mathematically, such affine transformations are of the form of $X^* = aX + b$, and $Y^* = cY + d$ [10]. Geometrical Moments (GM) of order $(p + q)$ for a continuous image function $f(x, y)$ are defined as :

$$M_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f(x, y)\, dx\, dy \tag{1}$$

where $p, q = 0, 1, 2, ..., \infty$. The above definition has the form of the projection of the function $f(x, y)$ onto the mononomial $x^p y^q$. In our case, where the function $f(x, y)$ has only two possible values of 0 and 1, the equation 1 reduces to :

$$M_{pq} = \sum_X \sum_Y x^p y^q f(x, y) \tag{2}$$

where X and Y represent x, y coordinates of the image. The center of gravity of the image has the coordinates :

$$\bar{x} = \frac{M_{10}}{M_{00}}, \bar{y} = \frac{M_{01}}{M_{00}}, \tag{3}$$

If we refer to the center of gravity as origin, we obtain :

$$\bar{M}_{pq} = \sum_X \sum_Y (x - \bar{x})^p (y - \bar{y})^q f(x, y) \tag{4}$$

These moments are also referred to as Central Moments and can be expressed as a linear combination of M_{jk} and the moments of lower order. The variances of the moment are defined as :

$$\sigma_x = \sqrt{\frac{\bar{M}_{20}}{M_{00}}}, \sigma_y = \sqrt{\frac{\bar{M}_{02}}{M_{00}}}, \tag{5}$$

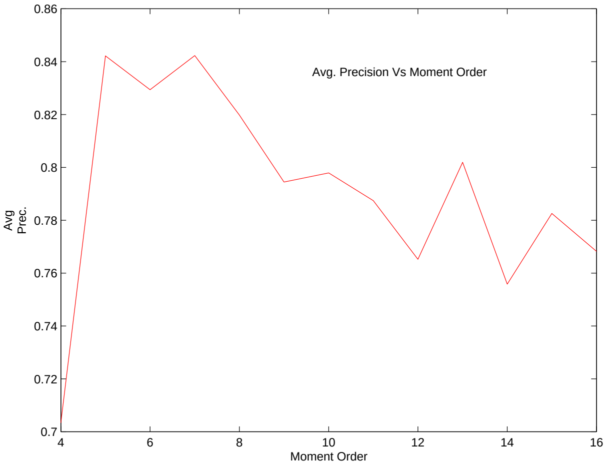


Figure 4: Average Precision curve Vs Moment Order for a Hindi Image Subset.

They are used to normalise the coordinates by setting:

$$x^* = \frac{(x - \bar{x})}{\sigma_x}, y^* = \frac{(y - \bar{y})}{\sigma_y}, \tag{6}$$

Using the normalised values of coordinates as obtained in equation 6 , the moment equation is as follows :

$$m_{pq} = \frac{\sum_X \sum_Y (x^*)^p (y^*)^q f(x, y)}{M_{00}} \tag{7}$$

which is invariant under image translation and scale transformations.

4 Feature Extraction and Indexing

Feature extraction is preceeded by preprocessing of documents prior to computing moment based functions. Firstly, the Horizontal Profile feature of the document image is used to segment into line images. Thereafter, Vertical Profile features of each line image is used to extract individual word images. The word images are normalised to equal height and width of 256 pixels.

Using equation 7, moments up to the 7th order are extracted from the normalised word images. A feature vector consisting of 30 moment values obtained is constructed for each word image and stored in the main index. Experiments were conducted to determine the number of orders up to which moments should be computed. As shown in Figure 4, average precision increases with the rise in moment orders (up to a threshold of 7 orders), after which the precision rate falls. This can be attributed to the nature of higher order Geometrical Moments which are prone to adding noise in the feature set and thereby reduce the overall precision after a certain threshold. After the index has been constructed using the moment features, we create a template which keeps the mapping between a word image and its corresponding text. This template is used to generate a query word image corresponding to the query text input by the user. A similar feature extraction mechanism is performed on the query word image to obtain a query feature vector which is used to find the similarity between the query word image and all other word images present in the corpus.

5 Similarity Matching and Relevance Feedback

5.1 Cosine Similarity

A standard Vector Space Model is used represent the query word and all the candidate words. The index is maintained in the form of a word-feature matrix, where each word image $\vec{w}$ occupies one row of the matrix and all columns in a single row correspond to the moment values computed for the given word image.

When the user enters any query word, a lookup operation is performed in the stored template to obtain the corresponding normalised word image for the input text. Feature extraction is performed on the word image to construct the query feature vector $\vec{q}$. A cosine similarity score is computed for this query feature vector and all the rows of the word-feature matrix. The cosine similarity is calculated as follows:

$$SIM(q,w) = \frac{\vec{q}\ .\ \vec{w}}{\mid \vec{q} \mid * \mid \vec{w} \mid} \tag{8}$$

All the words of the document corpus are then ranked according to the cosine similarity score. The top choice returned by the ranking mechanism represents the word image which is most similar to the input query word.

5.2 Relevance Feedback

Since the word images present in the document corpus may be of poor print quality and may contain noise, the moment features computed may not be effective in ranking relevant word images higher in the obtained result. Also the presence of higher order moments may lead to inconsistency in the overall ranking of word images. To overcome this limitation, we have implemented a Rocchio's formula based manual Relevance Feedback mechanism. This mechanism re-formulates the query feature vector by adjusting the values of the individual moment orders present in the query vector. The relevance feedback mechanism assumes a user input after the presentation of the initial results. A user enters either a 1 denoting a result to be relevant or 0 denoting a result to be irrelevant. The new query vector is computed as follows:

$$q_{new} = \gamma . q_{old} + \frac{\alpha}{|R|} . \sum_{i=1}^{i=R} d_i - \frac{\beta}{|NR|} . \sum_{j=1}^{j=NR} d_j \tag{9}$$

where α , β and γ are term re-weighting constants. R denotes a relevant result set and NR denotes a non-relevant result set. For this experiment, we chose $\alpha = 1$, $\beta = 0.75$ and $\gamma = 0.25$.

6 Experiments and Results

The moment based features seem more robust in handling different image transformations compared to commonly used feature sets for word spotting such as GSC features [1] and Gabor filter based features [2]. This can be obseerved in Figure 5. The first row of the image corresponds to different types of transformations applied to normal English handwritten word images ((a)) such as changing the image scale as in (b) or (c). The second row corresponds to linear ((f)) and scale transformation ((e)), when applied to the normal machine printed Hindi word image ((d)). Even after undergoing such transformations, the cosine similarity score between the moment features extracted from all image pairs is still close to 1, which reflects the strength of invariance of moment based features with respect to image transformations. Table 1 shows the cosine similarity score between all pairs of English word

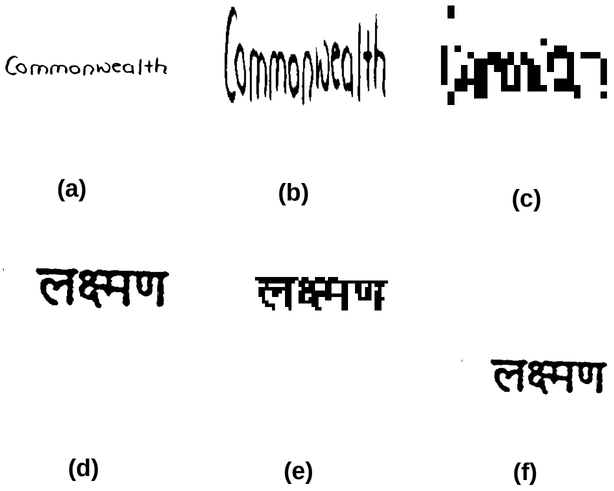


Figure 5: Various forms of Image Transformations. (a) & (d) Sample Word Image . (b),(c) & (e) Scale Transformation Examples (f) Linear Transformation Example .

Table 1: Cosine Similarity Score for English Transformed Word Image Pairs.

Word Image Pair	(a)	(b)	(c)
(a)	1	0.9867	0.9932
(b)	0.9867	1	0.9467
(c)	0.9932	0.9467	1

images. Table 2 shows the similarity score between all pairs of hindi word images.

The data set for evaluating our methodology consists of documents in three scripts, namely English, Hindi and Sanskrit. For English, we used publicly available IAMdb [13] handwritten word images and word images extracted from George Washington's publicly available historical manuscripts [14]. The dataset for English consists of 707 word images. For Hindi, 763 word images were extracted from publicly available Million Book Project documents [12]. For Sanskrit, 693 word images were extracted from 5 Sanskrit documents downloaded from the URL: http://sanskrit.gde.to/ . For public testing and evaluation, we have also made our dataset available at the location: http://cubs.buffalo.edu/ilt/dataset/.

For evaluating the system performance, we use the commonly used Average Precision Metric. Precision for

Table 2: Cosine Similarity Score for Hindi Transformed Word Image Pairs.

Word Image Pair	(d)	(e)	(f)
(d)	1	0.9662	0.9312
(e)	0.9662	1	0.9184
(f)	0.9312	0.9184	1

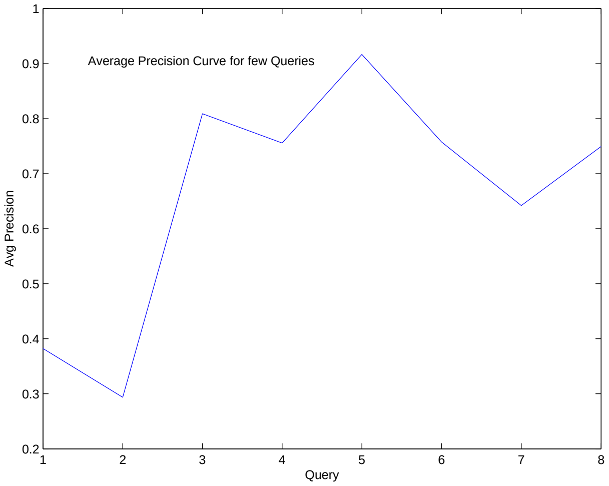


Figure 6: Average Precision curve for English Word Spotting.

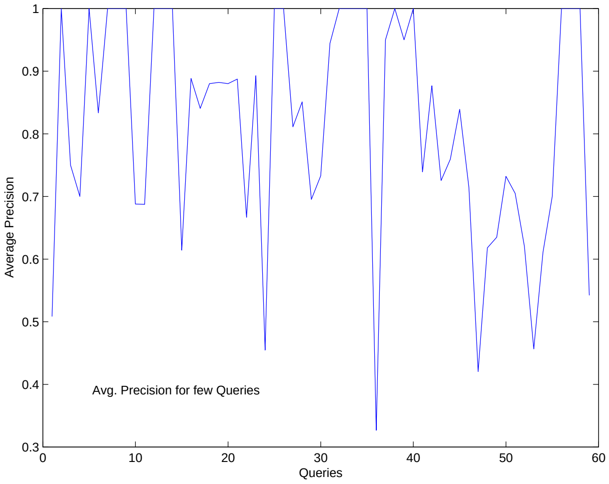


Figure 7: Average Precision curve for Hindi Word Spotting.

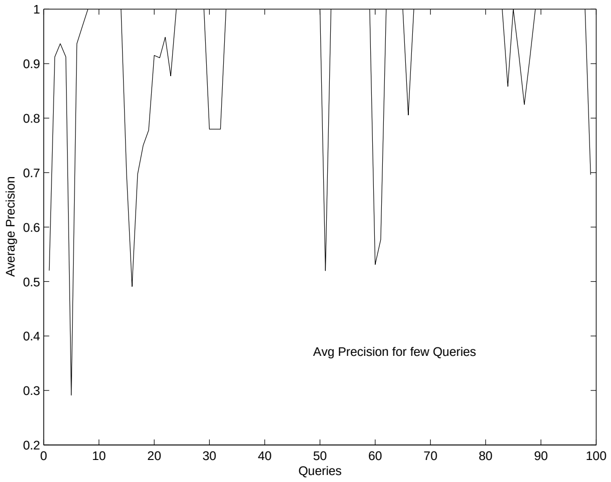


Figure 8: Average Precision curve for Sanskrit Word Spotting.

each query image was calculated at every recall level, and then averaged over to give an Average Precision per query. Figure 6 shows the average precision values for some query words in English. Figure 7 shows the average precision values for query words in Hindi. Figure 8 shows the average precision values for query words in Sanskrit.

The experimental results for all three scripts are summarised in Table 3. The Average Precision rates as shown in the table have been averaged over 30 queries in English, 75 queries in Hindi and 100 queries in Sanskrit. As shown here, the system works better for machine printed text (71.18 and 87.88) as compared to handwritten (67.0). The best performance is seen with Sanskrit script (87.88), which has a variable length words allowing it to be more discriminative in its feature analysis as compared to other two scripts. Table 4 compares the performance of GSC based word spotting as reported in [1] against our methodology. At 50% recall level, Moment based features perform better than GSC based features for both handwritten English and machine printed Sanskrit documents.

We also evaluate the performance of Gabor Feature based word spotting method [2] on our dataset. Features are extracted using an array of Gabor filters having a scale from 4 pixels to 6 pixels and 8 orientations. Table 5 summarizes the performance of Gabor features based method as opposed to our Moment based system. As shown , Moment based features outperform Gabor based features in terms of average precision rates obtained for all 3 scripts used in the experiment.

Table 3: Average Precision rate for word spotting in all 3 Scripts .

Script	Before RF	After RF
English	66.30	69.20
Hindi	71.18	74.34
Sanskrit	87.88	92.33

Table 4: Comparison of GSC and Moments based features at 50% recall level.

Script	GSC	Moments
English	60.0	71.6
Sanskrit	90.0	94.3

7 Summary and Conclusion

In this paper, we have proposed a framework for script independent word spotting in document images. We have shown the effectiveness of using statistical Moment based features as opposed to some of the structural and profile based features which may constrain the approach to few scripts. Another advantage of using moment based features is that they are image scale and translation invariant which makes them suitable for font independent feature analysis. In order to deal with the noise sensitivity of the higher order moments, we use a manual relevance feedback to improve the ranking of the relevant word images. We are currently working on extending our methodology to larger data sets and incorporating more scripts in future experiments.

Table 5: Comparison of Gabor filter based and Moments Features.

Script	Gabor	Moments
English	56.15	66.30
Hindi	67.25	71.18
Sanskrit	79.10	87.88

References

[1] S. N. Srihari, H. Srinivasan, C. Huang and S. Shetty, ”Spotting Words in Latin, Devanagari and Arabic Scripts,” Vivek: Indian Journal of Artificial Intelligence , 2006.

[2] Huaigu Cao, Venu Govindaraju, Template-Free Word Spotting in Low-Quality Manuscripts, the Sixth International Conference on Advances in Pattern Recognition (ICAPR), Calcutta, India, 2007.

[3] S. Kuo and O. Agazzi, Keyword spotting in poorly printed documents using 2-d hidden markov models, in IEEE Trans. Pattern Analysis and Machine Intelligence, 16, pp. 842848, 1994.

[4] M. Burl and P.Perona, Using hierarchical shape models to spot keywords in cursive handwriting, in IEEE-CS Conference on Computer Vision and Pattern Recognition, June 23-28, pp. 535540, 1998.

[5] A. Kolz, J. Alspector, M. Augusteijn, R. Carlson, and G. V. Popescu, A line oriented approach to word spotting in hand written documents, in Pattern Analysis and Applications, 2(3), pp. 153168, 2000.

[6] R. Manmatha and W. B. Croft, ”Word spotting: Indexing handwritten archives,”’ In M. Maybury, editor, Intelligent Multimedia Information Retrieval Collection , AAAI/MIT Press, Menlo Park, CA, 1997.

[7] T. Rath and R. Manmatha, .Features for word spotting in historical manuscripts,. in Proc. International Conference on Document Analysis and Recognition, pp. 218.222, 2003.

[8] T. Rath and R. Manmatha, .Word image matching using dynamic time warping,. in Proceeding of the Conference on Computer Vision and Pattern Recognition (CVPR), pp. 521.527, 2003.

[9] Teh C.-H. and Chin R.T., ”‘On Image Analysis by the Methods of Moments,”’ in IEEE Trans. Pattern Analysis and Machine Intelligence, 10, No. 4 , pp. 496513, 1988.

[10] Franz L. Alt , ”‘Digital Pattern Recognition by Moments,”’ in The Journal of the ACM , Vol. 9 , Issue 2 , pp. 240-258 , 1962.

[11] Jeff L. Decurtins and Edward C. Chen ,”’ Keyword spotting via word shape recognition ”’ in Proc. SPIE Vol. 2422, p. 270-277, Document Recognition II, Luc M. Vincent; Henry S. Baird; Eds. , vol. 2422 , pp. 270-277, March 1995.

[12] Carnegie Mellon University - Million book project, URL: http://tera-3.ul.cs.cmu.edu/, 2007.

[13] IAM Database for Off-line Cursive Handwritten Text, URL: http://www.iam.unibe.ch/ zimmerma/iamdb/iamdb.html .

[14] Word Image Dataset at CIIR - UMass , URL: http://ciir.cs.umass.edu/cgi-bin/downloads/downloads.cgi .