

# Pho(SC)Net: An Approach Towards Zero-shot Word Image Recognition in Historical Documents

Anuj Rai<sup>1</sup>, Narayanan C. Krishnan<sup>1</sup>, and Sukalpa Chanda<sup>2</sup>

<sup>1</sup> Department of Computer Science & Engineering,  
Indian Institute of Technology, Ropar, India  
{2019aim1003,ckn}@iitrpr.ac.in

<sup>2</sup> Department of Information Technology,  
Østfold University College, Norway  
sukalpa@ieee.org

**Abstract.** Annotating words in a historical document image archive for word image recognition purpose demands time and skilled human resource (like historians, paleographers). In a real-life scenario, obtaining sample images for all possible words is also not feasible. However, Zero-shot learning methods could aptly be used to recognize unseen/out-of-lexicon words in such historical document images. Based on previous state-of-the-art methods for word spotting and recognition, we propose a hybrid representation that considers the character’s shape appearance to differentiate between two different words and has shown to be more effective in recognizing unseen words. This representation has been termed as Pyramidal Histogram of Shapes (PHOS), derived from PHOC, which embeds information about the occurrence and position of characters in the word. Later, the two representations are combined and experiments were conducted to examine the effectiveness of an embedding that has properties of both PHOS and PHOC. Encouraging results were obtained on two publicly available historical document datasets and one synthetic handwritten dataset, which justifies the efficacy of “Phos” and the combined “Pho(SC)” representation.

**Keywords:** PHOS · Pho(SC) · Zero-shot word recognition · Historical Documents · Zero-shot learning · Word recognition.

## 1 Introduction

Historical documents could provide information and conditions about human societies in the past. Due to easy availability and usability of image acquisition devices such documents are being digitized and archived nowadays. Searching for important and relevant information from the large pool of images in the digital archive is a challenging task. Earlier, end-to-end transcription of the text using OCR was a popular way to achieve this goal. However, the performance of OCR often depends on character-segmentation accuracy, which is itself error prone,

specially in the context of cursive handwritten text. Moreover, end-users of such digital archives (historians, paleographers etc.) are often not interested in an end-to-end transcription of the text, rather they are interested in specific document pages where a query incident, place name, person name has been mentioned. To cater to this requirement, word-spotting and recognition techniques play an important role as they help directly in document indexing and retrieval.

Deep learning models have been quite successful in many document analysis problems. Thus, it is a natural fit for word recognition in historical documents as well. However, training deep networks for this problem is a challenging task due to many reasons. The lack of a large corpus of word images, partly due to the changes in the appearance of characters and spellings of words over centuries, makes it difficult to train a deep model. The complexity is further compounded by the requirement of learning large number of word labels, using only a small set of samples. In addition, the historical documents exhibit many undesirable characteristics such as torn and blurred segments, unwanted noise and faded regions, handwritten annotations by historians and artefacts; all of them contributing to the difficulty of the task.

Word spotting and recognition sound similar to each other but are two different tasks. This is illustrated in Figure 1. Given an image containing a word as input, a word recognition system identifies the word from the lexicon that is present in the image. On the other hand, word spotting refers to detecting other image segments in the document exhibiting patterns similar to the query image. This work is focused on the former problem and extends it to the zero-shot learning (ZSL) setting. Classical word recognition involves training a machine learning model to recognise the words given the images containing them. It is assumed that the test query images also contain only the words that were presented during the training. However, in the ZSL setting, the test query images can contain words that the model did not see during training. This is a more challenging task requiring a visual representation (akin to the semantic embedding in ZSL literature) that can bridge the set of seen and unseen words.

In this work, we propose a visual characterization of words to learn a mapping between word-images and their corresponding word labels such that it can also be used to recognise out-of-lexicon words. We call this characterization the Pyramidal Histogram of Shapes (PHOS). The PHOS representation encodes the primary shapes of character strokes of different word segments in a hierarchical manner. We use a deep convolutional network to learn the PHOS representation from images of words present in the training lexicon.

Overall we make the following contributions

- We present a novel representation of words (the PHOS representation) that encodes the visual features of the characters.
- The PHOS representation is used to perform zero-shot word recognition.
- Experiments on the PHOS and the popular PHOC representations suggest that PHOS encodes visual shape features of the characters that is missed by PHOC and therefore is more suitable to recognize unseen words.

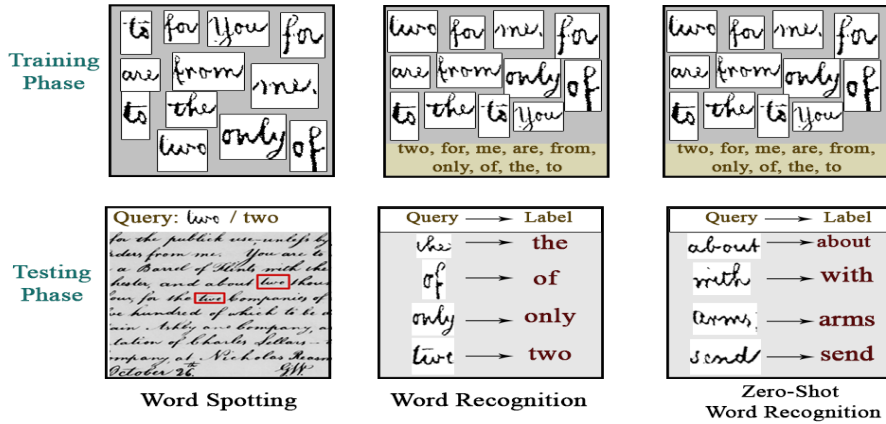


Fig. 1: Difference between word spotting and recognition: Given a word image as input, word recognition is finding the word in the lexicon that is the most likely to be the word present in the image. Word spotting involves finding image segments similar to the query image. Document image belongs to the GW dataset.

- Combining both PHOC and PHOS representations achieves the highest zero-shot word recognition accuracy on a synthetic and two real-world (George Washington and IAM Handwriting) datasets. However, for the generalized ZSL setting, the combined model is only able to outperform PHOS for the synthetic and IAM Handwriting datasets.

## 2 Related Work

Word spotting and recognition has been well explored over the last 25 years, with a spurt in deep learning based solutions in the recent past [20], [4], [21], [10], [6], [11], [7], [9], [12], [5], [22]. The seminal work [20] on word spotting involved training of a CNN for predicting the PHOC representation [2]. This work considered both contemporary as well as historical document images in their experiments. The proposed system can be used in both “Query By Example” (QBE) and “Query By String” (QBS) settings. In [11], the authors proposed an End2End embedding scheme to learn a common representation for word images and its labels, whereas in [7], a CNN-RNN hybrid model was proposed for handwritten word recognition. The method proposed in [4] uses a deep recurrent neural network (RNN) and a statistical character language model to attain high accuracy for word spotting and indexing. A recent work on robust learning of embeddings presents a generic deep CNN learning framework that includes pre-training with a large synthetic corpus and augmentation methods to generate real document images, achieving state-of-the-art performance for word spotting on handwritten and historical document images [12]. In [13], the authors introduce a novel semantic

representation for word images that uses linguistic and contextual information of words to perform semantic word spotting on Latin and Indic scripts. From the brief discussion it is evident that though deep learning based methods have been used in the recent past for word spotting and recognition tasks, zero-shot word recognition - recognizing a word image without having seen examples of the word during training; has not been studied. Only [6] explore Latin script word recognition problem in the ZSL framework; however the number of test classes were limited and no publicly available datasets were used in their experiments.

ZSL techniques have demonstrated impressive performances in the field of object recognition/detection [15], [25], [3],[17], [14],[24]. Li et.al. [15] propose an end-to-end model capable of learning latent discriminative features jointly in visual and semantic space. They note that user-defined signature attributes loose its discriminativeness in classification because they are not exhaustive even though they are semantically descriptive. Zhang et. al. [25], use a Graph Convolutional Network (GCN) along with semantic embeddings and the categorical relationships to train the classifiers. This approach takes as input semantic embeddings for each node (representing the visual characteristic). It predicts the visual classifier for each category after undergoing a series of graph convolutions. During training, the visual classifiers for a few categories are used for learning the GCN parameters. During the test phase, these filters are used to predict the visual classifiers of unseen categories [25]. In [3], the objective functions were customized to preserve the relations between the labels in the embedding space. In [1], attribute label embedding methods for zero-shot and few-shot learning systems were investigated. Later, a benchmark and systematical evaluation of zero-shot learning w.r.t. three aspects, i.e. methods, datasets and evaluation protocol was done in [23]. In [18], the authors propose a conditional generative model to learn latent representations of unseen classes using the generator trained for the seen classes. The synthetically generated features are used to train a classifier for predicting the labels of images from the unseen object category.

In summary, the current ZSL methods are focused towards object detection. There is no work on ZSL for word recognition. One must note that semantic attribute space in ZSL-based object detection is rich as attributes like colour and texture pattern play a crucial role. But in case of ZSL-based word recognition the semantic attribute space is rather constrained due to the absence of such rich visual features. Further the major bottleneck for ZSL based word recognition is the absence of the semantic embedding or attribute signature that establishes the relationship between the various word labels. This is further challenged by large number of word classes with relatively few examples. In this work, we propose a novel attribute signature and validate its effectiveness for ZSL-based word image recognition using standard benchmark datasets.

### 3 Methodology

We begin by defining our problem of interest. Let  $\mathcal{S} = \{(x_i, y_i, c(y_i))\}_{i=1}^N$ , where  $x_i \in \mathcal{X}, y_i \in \mathcal{Y}^s, c(y_i) \in \mathcal{C}$ ,  $\mathcal{S}$  stands for the training examples of seen word

labels,  $x_i$  is the image of the word,  $y_i$  is the corresponding word label in  $\mathcal{Y}^s = \{s_1, s_2, \dots, s_K\}$  consisting of  $K$  discrete seen word labels, and  $c(y_i) \in \mathcal{R}^Q$  is the unique word label embedding or attribute signature that models the visual relationship between the word labels. In addition, we have a disjoint word label set  $\mathcal{Y}^u = \{u_1, \dots, u_L\}$  of unseen labels, whose attribute signature set  $U = \{u_l, c(u_l)\}_{l=1}^L, c(u_l) \in \mathcal{C}$  is available, but the corresponding images are missing. Given  $\mathcal{S}$  and  $\mathcal{U}$ , the task of zero-shot word recognition is to learn a classifier  $f_{zsl} : \mathcal{X} \rightarrow \mathcal{Y}^u$  and in the generalized zero-shot word recognition, the objective is to learn the classifier  $f_{gzsl} : \mathcal{X} \rightarrow \mathcal{Y}^u \cup \mathcal{Y}^s$ . In the absence of training images from the unseen word labels, it is difficult to directly learn  $f_{zsl}$  and  $f_{gzsl}$ . Instead, we learn a mapping ( $\phi$ ) from the input image space  $\mathcal{X}$  to the attribute signature space  $\mathcal{C}$  that is shared between  $\mathcal{Y}^s$  and  $\mathcal{Y}^u$ . The word label for the test image  $x$  is obtained by performing a nearest neighbor search in the attribute signature space using  $\phi(x)$ . Thus, the critical features of zero-shot word recognition are the attribute signature space  $\mathcal{C}$  that acts as a bridge between the seen and unseen word labels and the mapping  $\phi$ . In this work, we propose a novel attribute signature representation that can effectively model the visual similarity between seen and unseen word labels. The mapping  $\phi$  is modeled as a deep neural network.

A very popular word label representation that can serve as the attribute signature for our problem is the pyramidal histogram of characters (PHOC). A PHOC is a pyramidal binary vector that contains information about the occurrence of characters in a segment of the word. It encodes the presence of a character in a certain split of the string representation of the word. The splits of different lengths result in the pyramidal representation. The PHOC allows to transfer information about the attributes of words present in the training set to the test set as long as all attributes in the test set are also present in the training set. However, this constraint may be violated in the context of zero-shot word recognition. Further, the PHOC also misses the visual shape features of the characters as they appear in a word image. We also observe these limitations of PHOC from our experiments on unseen word recognition. Literature also suggests parity between various representations that only encode the occurrence and position of characters within a word[19]. Thus we are motivated to present a novel attribute signature representation that complements the existing word label characterizations.

### 3.1 The Pyramidal Histogram of Shapes

We propose the pyramidal histogram of shapes (PHOS) as a robust bridge between seen and unseen words. Central to the PHOS representation is the assumption that every character can be described in terms of a set of primitive shape attributes [6]. We consider the following set of primitive shapes :- ascender, descender, left small semi-circle, right small semi-circle, left large semi-circle, right large semi-circle, circle, vertical line, diagonal line, diagonal line at a slope of 135 degrees, and horizontal line. These shapes are illustrated in Fig. 2. Only the counts of these shapes is insufficient to adequately characterize each word

uniquely. Inspired by the pyramidal capture of occurrence and position of characters in a word, we propose the pyramidal histogram of shapes that helps in characterizing each word uniquely.



Fig. 2: 11 primary shape attributes: ascender, descender, left small semi-circle, right small semi-circle, left large semi-circle, right large semi-circle, circle, vertical line, diagonal line, diagonal line at a slope of 135 degrees, and horizontal line

The process of capturing the PHOS representation for a word is illustrated in Fig.3. At the highest level of the pyramid, there exists only a single segment, which is the entire word. At every level  $h$  of the pyramid, we divide the word into  $h$  equal segments. Further, at every level  $h$ , we count the occurrence of the 11 primary shapes in every  $h$  segments of the word. The concatenation of the count vectors for every segment in a level and across all the levels of the pyramid results in the PHOS representation of the word. In this work, we have used levels 1 through 5, resulting in a PHOS vector of length  $(1+2+3+4+5)*11 = 165$ . Thus the PHOS vector encodes the occurrence and relative position of the shapes in the word string.

For example, let us consider a pair of anagrams “listen” and “silent” for 3 levels of segmentation. The segments at three levels for “listen” are:  $L_1 = \{listen\}$ ,  $L_2 = \{lis, ten\}$ ,  $L_3 = \{li, st, en\}$ . Similarly for “silent”:  $L_1 = \{silent\}$ ,  $L_2 = \{sil, ent\}$ ,  $L_3 = \{si, le, nt\}$ . The corresponding shape counts and their PHOS vector at each level for both words has been illustrated in Fig.3.

For the zero-shot word recognition problem, it is important to encode the occurrence and relative position of characters within a word, as well as that of visual shapes. Therefore, we propose to use the concatenated PHOC and PHOS vector of a word as its attribute signature representation  $\mathcal{C}$ . Thus, the attribute signature representation for the word label  $y_i$  is  $[c_c(y_i), c_s(y_i)]$ , where  $c_c(y_i)$  and  $c_s(y_i)$  are the PHOC and PHOS representations respectively.

### 3.2 Pho(SC)Net Architecture

Having defined the augmented attribute signature space  $\mathcal{C}$ , our next objective is to learn the mapping  $\phi$  to transform an input word image into its corresponding attribute signature representation - PHOC+PHOS vector. We use the architecture of SPP-PhocNet[20] as the backbone for the Pho(SC)Net ( $\phi$ ) that is used to predict the combined representation. The Pho(SC)Net is a multi-task network with shared feature extraction layers between the two tasks (PHOC and PHOS). The shared feature extraction network is a series of convolution layers,

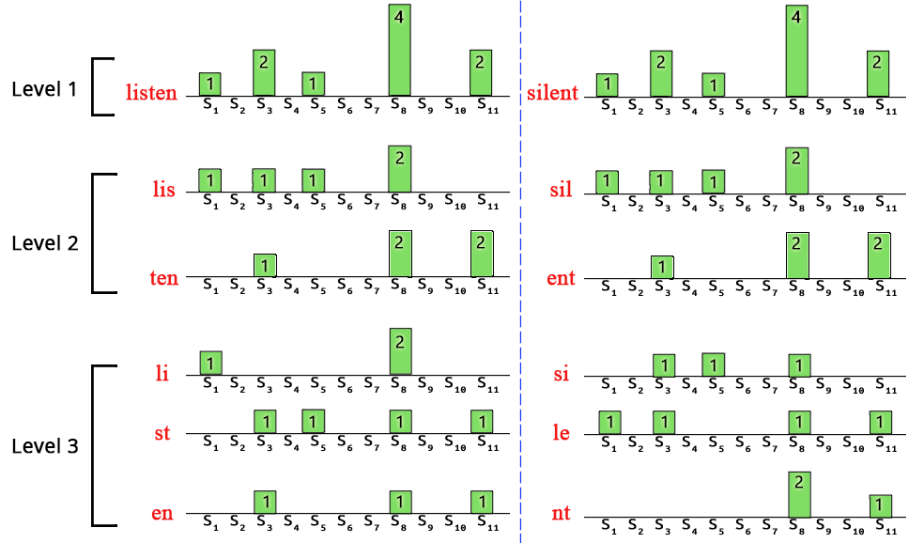


Fig. 3: Pyramidal structure of PHOS representation

followed by a spatial pyramid pooling (SPP) layer. The SPP layer facilitates the extraction of features across multiple image resolutions. The Pho(SC)Net separates out into two branches after the SPP layer to output the two representations. The two branches contain two independent fully connected layers. As the PHOC representation is a binary vector, the PHOC branch ends with a sigmoid activation layer. On the other hand the PHOS representation being a non-negative vector, the PHOS branch ends with a ReLU activation layer. The multi-task Pho(SC)Net architecture is illustrated in Fig.4.

The output of the Pho(SC)Net for an input word image is the vector  $\phi(x) = [\phi_C(x), \phi_S(x)]$ , where  $\phi_C(x)$  and  $\phi_S(x)$  are the predicted PHOC and PHOS representations respectively. Given a mini batch of  $B$  instances from the training set consisting of seen word images and their labels, we minimize the following loss function during training.

$$L = \sum_{i=1}^B \lambda_c L_c(\phi_c(x_i), c_c(y_i)) + \lambda_s L_s(\phi_s(x_i), c_s(y_i)) \quad (1)$$

where  $L_c(\phi_c(x_i), c_c(y_i))$  is the cross entropy loss between the predicted and actual PHOC representations,  $L_s(\phi_s(x_i), c_s(y_i))$  is the squared loss between the predicted and actual PHOS representations, and  $\lambda_c, \lambda_s$  are hyper-parameters used to balance the contribution of the two loss functions.

Given a test image  $x$ , the Pho(SC)Net is used to predict the PHOC and PHOS representations to obtain the predicted attribute signature representation  $[\phi_c(x), \phi_s(x)]$ . The word whose attribute signature representation has the highest similarity (measured as cosine similarity) with  $[\phi_c(x), \phi_s(x)]$  is the predicted

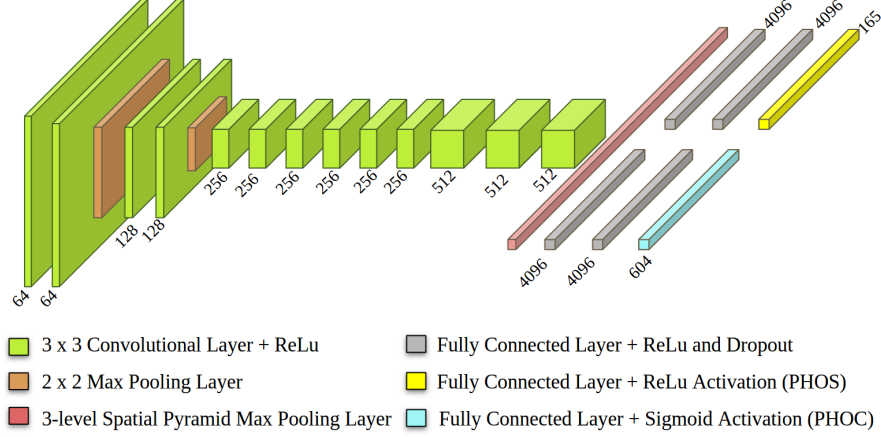


Fig. 4: Architecture of the multi-task Pho(SC)Net

word label for the test image, in the conventional ZSL setting, as defined below

$$\hat{y} = \operatorname{argmax}_{k \in \mathcal{Y}^u} \cos([\phi_c(x), \phi_s(x)]^T [c_c(k), c_s(k)]) \quad (2)$$

## 4 Experiments

### 4.1 Datasets

We validate the effectiveness of the Pho(SC) representation for the zero-shot word recognition problem on the following three datasets.

**Most Frequently Used Words (MFU) Dataset** A synthetic dataset was created from the most frequently used English words list.  $\mathcal{Y}^s$  was chosen to be the first 2000 words and the subsequent 1000 words were made part of  $\mathcal{Y}^u$ . Eight handwriting fonts were used to generate a total of 16000-word images (split into 12000 for training and 4000 for testing) for  $\mathcal{Y}^s$  and 8000 for  $\mathcal{Y}^u$ . We ensure that the word labels of the 4000 test images of  $\mathcal{Y}^s$  are present in the corresponding training set. These details are summarized in Table 1.

**George Washington (GW) Dataset** The George Washington dataset[8] exhibits homogeneous writing style and contains 4894 images of 1471 word labels. We used the lower-case word images from the standard four-fold cross-validation splits accompanying the dataset to evaluate the Pho(SC)Net. We modified the validation and test sets to suit the zero-shot word recognition problem. Specifically, the test sets in each split was further divided into two parts: seen and unseen word label images, and the validation set contained only seen word label images. The details for each set across all the splits are presented in Table 1.



**IAM Handwriting (IAM) Dataset** IAM handwriting dataset[16] is a multi-writer dataset that consists of 115320 word-images, from 657 different writers. We used the lowercase word-class images from this dataset to create train, validation, and test sets. Specifically we created two different splits. In the first split (overlapping writers ZSL split), we ensured that the writers in the test set are also part of the training set. Further, the test set contained both unseen and seen word labels, while the train and validation sets contained only seen word images. In the second split (standard ZSL split) derived from the standard split accompanying the dataset we removed the unseen word images from the validation set, and divided the test split further into seen word and unseen word images. The standard ZSL split has an additional challenge as the writers in the training, validation, and test sets are non-overlapping. The number of images and the (seen and unseen) word labels for each split is presented in Table 1.













George Washington Dataset	 they	 hundred	 delivered	 publick
IAM Dataset	 ages	 forward	 transcendent	 through
MFU Synthetic Dataset	 than	 rightward	 margins	 agile

Fig. 5: Examples of word images and their labels from the three datasets

The training set of all the three datasets were further augmented through shearing, and addition of Gaussian noise. The size of a word image (in the training, validation, and test sets) depends on the length of the word and the handwriting style (font), but we needed images of uniform size for training. Hence, the binarized images were resized to the best fitting sizes (without changing the aspect ratio) and then padded with white pixels to get images of size 250\*50. Fig 5 presents a few examples of word images from the three datasets.

## 4.2 Training and Baselines

The Pho(SC)Net was trained using an Adam optimizer with learning rate 1e-4, weight decay set as 5e-5, momentum at 0.9. The batch size is kept as 16. The hyper-parameters  $\lambda_c$  and  $\lambda_s$  were fine tuned using the validation set. The final values chosen for these parameters are 1 and 4.5 respectively. We also conducted ablation studies with the individual PHOCNet (SPP-PHOCNet) and the PHOSNet counter part. These two networks were also trained in a similar fashion for all the three datasets. This helps to investigate the effect of adding visual shape representations to the default attribute signature vector

Split	Train Set	Validation Set	Test (Seen Classes)	Test (Unseen Classes)
<b>MFU Dataset</b>				
<b>MFU 2000</b>	36000(2000)	3600	4000	8000(1000)
<b>George Washington (GW Dataset)</b>				
<b>Split 1</b>	1585(374)	662(147)	628(155)	121(110)
<b>Split 2</b>	1657(442)	637(165)	699(155)	114(100)
<b>Split 3</b>	1634(453)	709(164)	667(148)	105(89)
<b>Split 4</b>	1562(396)	668(149)	697(163)	188(152)
<b>IAM Handwriting Dataset</b>				
<b>ZSL Split</b>	30414(7898)	2500(1326)	1108(748)	538(509)
<b>Standard Split</b>	34549(5073)	9066(1499)	8318(1355)	1341(1071)

Table 1: Details of dataset used for experiments *Numbers inside the parentheses represent the number of word classes in the set.*

(PHOC vector). Early stopping was applied to the training process using reducedLR on plateau on the validation set. As this is the first work to perform zero-shot word recognition on publicly available benchmark datasets, we compare Pho(SC)Net with the classical SPP-PHOCNet model (trained using the settings in[20]). An implementation of the proposed method can be found in [github.com/anuj-rai-23/PHOSC-Zero-Shot-Word-Recognition](https://github.com/anuj-rai-23/PHOSC-Zero-Shot-Word-Recognition).

### 4.3 Performance Metrics

The top-1 accuracy of the model’s prediction is used as the performance metric for all the experiments. The top-1 accuracy measures the proportion of test instances whose predicted attribute signature vector is closest to the true attribute signature vector. At test time, in the ZSL setting, the aim is to assign an unseen class label, i.e.  $\mathcal{Y}^u$  to the test word image and in the generalized ZSL setting (GZSL), the search space includes both seen and unseen word labels i.e.  $\mathcal{Y}^u \cup \mathcal{Y}^s$ . Therefore, in the ZSL setting, we estimate the top-1 accuracy over  $\mathcal{Y}^u$ . In the GZSL setting, we determine the top-1 accuracy for both  $\mathcal{Y}^u$  and  $\mathcal{Y}^s$  independently and then compute their harmonic mean.

## 5 Results and Discussion

The accuracies for the unseen word labels under the conventional ZSL setting is presented in Table 2. Overall, it is observed that the PHOC representation is not well-suited for predicting unseen word labels. However, the PHOS representation is more accurate in predicting the unseen word labels. Further, the combination of both the vectors (Pho(SC)) results in a significant improvement (on an average  $> 5\%$ ) in the unseen word prediction accuracy. The MFU dataset, on the account of being synthetically generated and noise-free, has the highest unseen word recognition accuracy. We obtained the least accuracy on the GW dataset split 4. We attribute this low accuracy to a rather large number of unseen word classes in the test set for this split.

Split	PHOC	PHOS	Pho(SC)
<b>MFU Dataset</b>			
MFU 2000	.94	.96	<b>.98</b>
<b>GW Dataset</b>			
GW Split 1	.46	.61	<b>.68</b>
GW Split 2	.64	.72	<b>.79</b>
GW Split 3	.65	.71	<b>.80</b>
GW Split 4	.35	<b>.62</b>	.60
<b>IAM Handwriting Dataset</b>			
ZSL Split	.78	.79	<b>.86</b>
Standard Split	.89	.88	<b>.93</b>

Table 2: ZSL Accuracy on all the splits

It is also observed that the accuracy of the model on MFU and IAM datasets are significantly higher than that of any of the GW splits. This is explained by observing the number of seen classes the model is presented during training. Both MFU and IAM datasets have more than 2000 seen word labels, allowing the model to learn the rich relationships between the word labels as encoded by the attribute signatures. Learning this relationship is essential for the model to perform well on unseen word images.

We also observe that using only PHOS as the attribute signature results in better performance than using PHOC on datasets that have homogeneous writing style. This is inferred by noticing the significant increase in the unseen word accuracy of over 14% on the GW dataset splits by PHOS over PHOC.

Split	PHOC			PHOS			Pho(SC)		
	$A_u$	$A_s$	$h$	$A_u$	$A_s$	$h$	$A_u$	$A_s$	$h$
<b>MFU Dataset</b>									
MFU 2000	.74	.99	.85	.92	.93	.92	.92	.99	<b>.96</b>
<b>GW Dataset</b>									
GW Split 1	.01	.96	.03	.24	.95	<b>.39</b>	.15	.97	.27
GW Split 2	.10	.98	.19	.30	.98	.46	.30	.98	<b>.46</b>
GW Split 3	.09	.97	.17	.40	.94	<b>.56</b>	.35	.96	.51
GW Split 4	.04	.94	.08	.31	.92	<b>.47</b>	.25	.95	.39
<b>IAM Dataset</b>									
ZSL Split	.58	.88	.70	.71	.82	.76	.77	.93	<b>.84</b>
Standard Split	.46	.87	.61	.64	.82	.72	.70	.90	<b>.79</b>

Table 3: Generalized ZSL accuracy on various splits.  $A_u$  = Accuracy with unseen word classes,  $A_s$  = Accuracy with seen word classes, Generalized ZSL accuracy,  $h$  = Harmonic mean of  $A_u$  and  $A_s$ .

Correctly classified samples of seen words					
something	man	moment	career	knows	distinguished
course	water	which	gin	knew	openend
Correctly classified samples of unseen words					
debauchery	helicopter	instinctively	irresolute	flattish	expecting
slugged	nonplussed	stranger	submariners	repressed	circling
Misclassified samples of seen words as seen words					
eight night	often offer	round sound	describe distinguished	gate cigarette	what combination
Misclassified samples of seen words as unseen words					
ground grand	climbed chimed	precious repercussions	harm ham	than threatening	won mentioned
Misclassified samples of unseen words as seen words					
bus ten	snub smell	crow now	obsession decision	jetty specially	clockwork color
Misclassified samples of unseen words as unseen words					
trail fail	courteously continuously	retrieved relieved	closer secrets	dockyard destroyed	whiskers unrecognisable
Green = True Label    Crimson = Predicted Incorrect Label    Blue = Predicted Correct Label					

Fig. 6: Examples of correct and incorrect predictions in Generalized ZSL setting from the standard IAM Split

Table 3 presents the test seen and unseen word accuracies, along with the harmonic mean for the GZSL setting. The high accuracies on the seen word labels and low accuracies on unseen word labels for the model using only PHOC seems to suggest that the PHOC representation is more suitable for scenarios where the word labels across the train and test are similar. In contrast, the PHOS model has marginally lower accuracies on the test seen word labels (in comparison to PHOC), but significantly higher accuracies on the unseen word labels (again in comparison to PHOC). This indicates that looking at visual shapes is more reliable when the train and test sets contain different word labels. Further, we also observe that the combined model yields higher performance for both MFU and IAM datasets that have a large number of seen classes. However, on the GW dataset, that has a significantly smaller number of seen word classes, the PHOS model is able to even outperform the combined model. The poor performance of the PHOC model that is pulling down the performance of the combined model on the unseen classes for the GW dataset can also be attributed to the small number of seen word classes that the model is exposed to during training. However, note that the PHOC model achieves significantly higher accuracies on the test images

of the seen word classes, indicating that the model has not overfit, but is biased towards these classes.

**Error Analysis** Fig. 6 illustrates a few predictions by the Pho(SC) model on the IAM dataset under the GZSL setting. It can be observed that even when the model incorrectly predicts the word, there is a good overlap between the characters of the predicted and true word label.

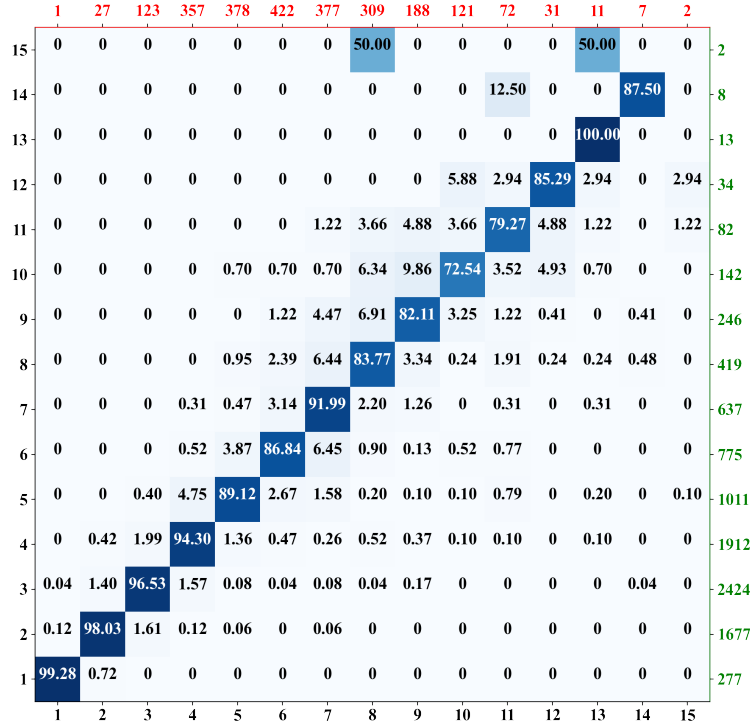


Fig. 7: Lengthwise confusion matrix(normalised) for predicted word length(left axis) and true word length(bottom) for Standard IAM Split. Labels on top(red) indicate the number of word classes in lexicon of that length while on right(green) represent the number of samples in test set for the corresponding word length.

Fig. 7 presents the confusion matrix for predictions on the IAM dataset standard split in the GZSL setting. The confusion matrix has been computed between words of different lengths to uncover any biases of the model (if any). It is difficult to visualize the class specific confusion matrix as there are over 1000 word labels with very few (often only 1) images per word label. The length of the predicted word labels is mostly within a range of the length of the true word

label. In general, the model is not biased towards words of any specific length. The high values along the diagonal indicates that the model is often predicting the word of the correct length (except when the word length is 15, for which there are only 2 samples and 2 classes).

## 6 Conclusion and Future Work

In this paper, we present the first exhaustive study on zero-shot word recognition for historical document images. We propose a novel attribute signature representation (PHOS) that characterizes the occurrence and position of elementary visual shapes in a word. Our experiments demonstrate the effectiveness of PHOS for predicting unseen word labels in the ZSL setting, while the classical PHOC representation is more suitable for seen word labels. Further, we combine both the representations to train a multi-task model- Pho(SC)Net that achieves superior performance over the individual representations. We validate the performance of the models on standard benchmark datasets as well as on a synthetically generated handwritten words dataset. Future directions to this work include extending the PHOS representations to include other non-Latin scripts like Arabic, Chinese, and Indic scripts.

## References

1. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(7), 1425–1438 (2016)
2. Almazán, J., Gordo, A., Fornés, A., Valveny, E.: Word spotting and recognition with embedded attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(12), 2552–2566 (2014)
3. Annadani, Y., Biswas, S.: Preserving semantic relations for zero-shot learning. In: *The IEEE Conference on Computer Vision and Pattern Recognition* (2018)
4. Bluche, T., Hamel, S., Kermorvant, C., Puigcerver, J., Stutzmann, D., Toselli, A.H., Vidal, E.: Preparatory KWS experiments for large-scale indexing of a vast medieval manuscript collection in the HIMANIS project. In: *International Conference on Document Analysis and Recognition*. pp. 311–316 (2017)
5. Carbonell, M., Fornés, A., Villegas, M., Lladós, J.: A neural model for text localization, transcription and named entity recognition in full pages. *Pattern Recognition Letters* **136**, 219–227 (2020)
6. Chanda, S., Baas, J., Haitink, D., Hamel, S., Stutzmann, D., Schomaker, L.: Zero-shot learning based approach for medieval word recognition using deep-learned features. In: *International Conference on Frontiers of Handwriting Recognition*. pp. 345–350 (2018)
7. Dutta, K., Krishnan, P., Mathew, M., Jawahar, C.V.: Improving CNN-RNN hybrid networks for handwriting recognition. In: *International Conference on Frontiers of Handwriting Recognition*. pp. 80–85 (2018)
8. Fischer, A., Keller, A., Frinken, V., Bunke, H.: Lexicon-free handwritten word spotting using character hmms. *Pattern recognition letters* **33**(7), 934–942 (2012)

9. Graves, A., Schmidhuber, J.: Offline handwriting recognition with multidimensional recurrent neural networks. In: *Advances in Neural Information Processing Systems*. pp. 545–552 (2009)
10. Kang, L., Toledo, J.I., Riba, P., Villegas, M., Fornés, A., Rusiñol, M.: Convolve, attend and spell: An attention-based sequence-to-sequence model for handwritten word recognition. In: *German Conference on Pattern Recognition*. pp. 459–472 (2018)
11. Krishnan, P., Dutta, K., Jawahar, C.V.: Word spotting and recognition using deep embedding. In: *Document Analysis Systems*. pp. 1–6 (2018)
12. Krishnan, P., Jawahar, C.: Hwnet v2: An efficient word image representation for handwritten documents. *International Journal on Document Analysis and Recognition* **22**(4), 387–405 (2019)
13. Krishnan, P., Jawahar, C.: Bringing semantics into word image representation. *Pattern Recognition* **108** (2020)
14. Li, K., Min, M.R., Fu, Y.: Rethinking zero-shot learning: A conditional visual classification perspective. In: *IEEE International Conference on Computer Vision*. pp. 3583–3592 (October 2019)
15. Li, Y., Zhang, J., Zhang, J., Huang, K.: Discriminative learning of latent features for zero-shot recognition. In: *The IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7463–7471 (2018)
16. Marti, U.V., Bunke, H.: The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition* **5**(1), 39–46 (2002)
17. Niu, L., Veeraraghavan, A., Sabharwal, A.: Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2018)
18. Paul, A., Krishnan, N.C., Munjal, P.: Semantically aligned bias reducing zero shot learning. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7056–7065 (2019)
19. Sudholt, S., Fink, G.A.: Evaluating word string embeddings and loss functions for cnn-based word spotting. In: *International Conference on Document Analysis and Recognition*. pp. 493–498 (2017)
20. Sudholt, S., Fink, G.A.: Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In: *15th International Conference on Frontiers in Handwriting Recognition, 2016*. pp. 277–282 (2016)
21. Wilkinson, T., Lindström, J., Brun, A.: Neural ctrl-f: Segmentation-free query-by-string word spotting in handwritten manuscript collections. In: *International Conference on Computer Vision*. pp. 4443–4452 (2017)
22. Wolf, F., Fink, G.A.: Annotation-free learning of deep representations for word spotting using synthetic data and self labeling. In: *Document Analysis Systems. Lecture Notes in Computer Science*, vol. 12116, pp. 293–308. Springer (2020)
23. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning - the good, the bad and the ugly. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3077–3086 (2017)
24. Xie, G.S., Liu, L., Jin, X., Zhu, F., Zhang, Z., Qin, J., Yao, Y., Shao, L.: Attentive region embedding network for zero-shot learning. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2019)
25. Zhang, H., Koniusz, P.: Zero-shot kernel learning. In: *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7670–7679 (2018)