# Robust Output Modeling in Bag-of-Features HMMs for Handwriting Recognition

Leonard Rothacker
*Depertment of Computer Science*
*TU Dortmund University*
*Dortmund, Germany*
*leonard.rothacker@udo.edu*

Gernot A. Fink
*Depertment of Computer Science*
*TU Dortmund University*
*Dortmund, Germany*
*gernot.fink@udo.edu*

***Abstract*—Bag-of-Features HMMs have been successfully applied to handwriting recognition and word spotting. In this paper we extend our previous work and present methods for modeling sequences of Bag-of-Features representations with Hidden Markov Models. We will discuss our previous approach that uses a pseudo-discrete model. Afterwards, we present a novel semi-continuous integration. The method is effective for probabilistic text clustering and is suitable for statistically modeling the characteristics of Bag-of-Features representations extracted from document images. Furthermore, its statistical expectation-maximization estimation can directly be integrated in Baum-Welch HMM training. In our experiments we present competitive results on the IfN/ENIT word recognition benchmark and state-of-the-art results for word spotting on the George Washington benchmark. Our evaluation gives insights into the properties of the models from the perspectives of modern as well as historic document analysis.**

***Keywords*-Bag-of-Features; Hidden Markov Models; handwriting recognition; word spotting;**

## I. Introduction

Bag-of-Features (BoF) have successfully been applied for representing natural scene images, cf. [1], as well as document images, cf. [2]. Their ability to automatically adapt themselves to the problem domain in an unsupervised manner is very powerful and leads to very high recognition rates in relation to the simplicity of the model [3]. Especially with respect to historic document images, the application of BoF representations leads to considerable advances. As the characteristics of historic documents change rapidly from document collection to document collection, an automatic adaptation is very desirable [4]. By modeling sequences of BoF representations with Hidden Markov Models (HMMs), we combine these positive properties with the advantages of modeling the sequential structure of handwritten script for word recognition [5] and word spotting [6].

The main idea for BoF is derived from machine-readable text categorization and retrieval where texts are represented as histograms of term frequencies, so called Bag-of-Words (BoW). In this model terms are words that are typical for the problem domain. Since all texts are represented by the same set of terms, it is referred to as vocabulary. BoF can be defined in analogy by replacing the words from the text domain with a suitable feature from the target domain. For images, the SIFT descriptor [7], also cf. [1], has been most successful. In order to find descriptors that are typical for the problem domain, clustering is applied. In analogy to BoW, the cluster representatives are called visual words in a visual vocabulary. In practice, BoW and BoF representations are very high dimensional and sparse.

Since BoF representations are a direct generalization of BoW, most of the related text processing methodology is directly applicable to BoF as well. For image retrieval, images from databases and respective image queries can be represented as BoF and processed with common text retrieval methods [8]. Statistical models for BoW representations, also in conjunction with low-dimensional vector space embeddings, so-called topic models, have been studied extensively for text and image processing, cf. e.g., [9], [10].

In order to represent BoF in HMMs, they have to be modeled as observations in the statistical HMM process. However, due to their discrete and very high dimensional nature, standard approaches, like continuous and also semi-continuous Gaussian mixture model (GMM) integrations, are not directly applicable. Reducing the dimensionality of the features would be required first, cf. [11] Chap. 9.

The main contribution of this paper is a discussion and evaluation of two different output modeling techniques for BoF representations in HMMs. We give insights to our previous approach [5] and we present a novel HMM integration that to the best of our knowledge has not been considered before. This model is adapted to the special data characteristics and, therefore, allows for a robust and integrated estimation in Baum-Welch HMM training. No separate dimensionality reduction model is required. Avoiding separate dimensionality reduction is an advantage, also because it has shown inferior handwriting recognition performance compared to our more direct integration in [5].

In Section II we will discuss related statistical models and HMM integrations. Afterwards, we will present BoF-HMMs for an application in handwriting recognition (Section III). In this regard, we will explain the statistical properties of our BoF representations along with two different output models. Our discussion will be supported by an experimental evaluation in Section IV before summarizing our results in Section V.

2167-6445/16 $31.00 © 2016 IEEE
DOI 10.1109/ICFHR.2016.44

## II. Related Work

HMMs model the generation of observation sequences in a two stage stochastic process, cf. [11] Chap. 5. In the first stage the probabilistic transitions over a finite state space are modeled. In the second stage observations are generated for each point in time within the process. For recognition, states are associated with classes and the state sequence that generated the observation sequence most likely, is decoded. In our case the generation of observations is of special interest. The probability / density for an observation $\mathbf{x}_t$ in state $j$ at time $t$ is given as

$$b_j(\mathbf{x}_t) = p(\mathbf{x}_t | S_t = j). \tag{1}$$

The observations $\mathbf{x}_t = (x_{t1}, \cdots, x_{tD})^\top$ are $D$-dimensional feature vectors and the function $b_j$ refers to a suitable discrete probability mass function or continuous probability density function. For applications of HMMs in handwriting recognition and word spotting, $b_j$ is often defined as a GMM.

Rodríguez-Serrano and Perronnin [12] model handwritten word images with semi-continuous Gaussian mixture HMMs for word spotting. Since in the semi-continuous case the Gaussian densities are shared among all states, the authors have an interesting interpretation related to BoF representations. They consider the shared codebook as a visual vocabulary and the probabilistic assignment of a feature vector to density components in the mixture model as a soft BoF representation.

In our case, however, we extract the BoF representations directly from word images. For this reason, GMMs cannot be estimated easily due to the very high dimensionality of BoF vectors. Furthermore, they do not fit well with their statistic properties. In the following we will present statistical models that have been specifically chosen in this regard.

Giménez et al. [13] apply Bernoulli HMMs to handwriting recognition. Due to its application domain and usage of a specific output model it is highly related to our work. For recognition they binarize word images and extract sequences of binary vectors in a sliding window manner. These are considered as observations. Their HMMs model these observations $\mathbf{x}_t \in \{0,1\}^D$ with $M$ mixtures of Bernoulli distributions weighted by mixture coefficients $c_{jk}$:

$$b_j(\mathbf{x}_t) = \sum_{k=1}^{M} c_{jk} \prod_{d=1}^{D} p_{jkd}^{x_{td}} (1 - p_{jkd})^{1-x_{td}}. \tag{2}$$

The mixture components are defined as the product of $D$ independent Bernoulli probability functions. The parameters $p_{jkd}$ indicate the probability of generating a black pen stroke pixel for feature component $d$ in mixture component $k$ for state $j$. Therefore, the distribution fits well with the characteristics of the feature vectors. Different mixture components allow for modeling different character shapes in the same model. Please note that mixture models require a convex linear combination of mixture components. Parameter estimates can be computed in an expectation-maximization (EM) fashion which is integrated in Baum-Welch HMM training.

Frasconi et al. [14] present a method for multi-page text categorization with HMMs. Individual pages are represented as BoW which makes their HMM integration highly related to our work. Using the HMM they are able to exploit internal document structure, like *preface*, *table of contents*, *main text* or *index*. This structure can either be hand-crafted or inferred automatically by state clustering. In order to model BoW representations as output of the HMM, they use multinomial distributions. Observations $\mathbf{x}_t \in \mathbb{N}_0^V$ are therefore count vectors of term occurrences in page $t$. $V$ indicates the size of the vocabulary, i.e., the number of terms $w_v$. For modeling the document structure, the probability mass functions are not state, but class-dependent, i.e., state $j$ is associated with class $C_\kappa$ which is indicated by $\phi$:

$$b_j(\mathbf{x}_t) = \prod_{v=1}^{V} p(w_v | S_t = j, \phi(j) = C_\kappa)^{x_{tv}}. \tag{3}$$

In the multinomial distribution, $p(w_v | S_t = j, \phi(j) = C_\kappa)$ is the probability for term $w_v$ in class $C_\kappa$. By weighting these probabilities in the exponent, the overall product is influenced by the class-specific term probabilities according to their occurrence in the page. It is to be noted that Equation 3 does not contain the multinomial coefficient, e.g., as defined in [15]. In practice, both variants of the multinomial distribution can be found. Parameters are estimated by computing the ratio of occurrences of a specific term in a given class and the number of occurrences of all terms in that class, where $N(w_v, C_\kappa)$ is the number of occurrences of term $w_v$ in class $C_\kappa$:

$$p(w_v | C_\kappa) = \frac{1 + N(w_v, C_\kappa)}{V + \sum_{r=1}^{V} N(w_r, C_\kappa)}. \tag{4}$$

It is important to note that neither of the term probabilities should be zero. This would be the case if a specific term does not occur in a specific class. Therefore, Laplacian smoothing is applied for regularization in Equation 4.

For our approach of extracting sequences of BoF representations from text images, this last aspect is particularly important. Laplacian smoothing works well if terms occur within all classes most of the time. Typically this is the case for densely populated Bag-of representations. For example, in [14] BoW representations are extracted on class level rather then document or page level for parameter estimation, i.e., the multinomials are class- and not HMM state-dependent. However, in our case BoF representations are extremely sparse.

Comparable scenarios with BoW representations can be found for short text clustering, e.g., twitter message analysis [16]. In the text processing literature these are approached, for example, with advanced multinomial models [16], [17].

## III. Output Modeling in Bag-of-Features HMMs

BoF-HMMs model BoF sequences that are typical for characters or words. Figure 1 shows the overall process for recognizing word images. In Section III-A we will outline how BoF sequences are extracted. Section III-B describes two different approaches for BoF output modeling.

### A. Bag-of-Features Sequences

In order to represent document images with BoF, SIFT descriptors [7] are computed in a dense grid. For word recognition we choose a descriptor cell structure of 4 rows and 2 columns instead of the 4×4 standard. This rectangular shape is better suited when using character models as less horizontal context is captured. The dense grid resolution depends on the resolution and size of the descriptors. Here, we extract descriptors in a 5×5 grid. An interesting aspect is the pruning of descriptors that do not contain any relevant information. A simple heuristic is to discard descriptors in low contrast regions, as image regions containing pen strokes have high contrast. We use descriptor pruning in order to demonstrate robustness of our output models in Section IV. An example for descriptor pruning is also shown in Figure 1.

In the next step, descriptors are quantized with respect to a pre-computed visual vocabulary. The visual vocabulary is obtained by clustering descriptors from a training set with Lloyd's algorithm, cf. [11] Sec. 4.3.1. For handwriting recognition and word spotting large vocabularies of a few thousand visual words give good results, cf. [5], [4]. The sequence of BoF representations is then obtained by sliding a window over the grid columns in writing direction.

### B. Output Modeling

For modeling the generation of BoF representations in the HMM we present two different approaches. The first approach has been presented for Arabic handwriting recognition [5] and has also been applied to word spotting, cf. [6]. It directly models visual word probabilities within HMM states (Section III-B1). The second approach is a novel HMM output model integration which is inspired by probabilistic short text clustering. The model was presented in [17] and can directly be integrated in a semi-continuous HMM (Section III-B2).

*1) Pseudo-Discrete Model:* This BoF output model can be interpreted as a soft extension of a discrete model. For that reason we referred to it as pseudo-discrete, cf. [5]. It allows for modeling the observation of multiple visual words at a point in time. For relative visual word frequencies $\mathbf{x}_t \in \mathbb{Q}^V_{\geq 0}$ we obtain the probability for generating a BoF representation at time $t$ in state $j$:

$$b_j(\mathbf{x}_t) = \sum_{v=1}^{V} p(w_v|S_t = j)x_{tv} \ , \quad \sum_{v=1}^{V} x_{tv} = 1. \tag{5}$$

Given training data $\mathbf{O} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$ and the current model $\lambda$, parameters are estimated using the probability $p(W_t = w_v|\mathbf{O}, \lambda)$ of observing visual word $w_v$ at time $t$ and probability $p(S_t = j|\mathbf{O}, \lambda)$ of state $j$ being active at time $t$:

$$p(w_v|S_t = j) = \frac{\sum_{t=1}^{T} p(W_t = w_v, S_t = j|\mathbf{O}, \lambda)}{\sum_{t=1}^{T} p(S_t = j|\mathbf{O}, \lambda)}. \tag{6}$$

The major difference to the multinomial model, cf. Equation 3, is that it does not model a specific configuration of independent observations but *any* configuration of independent observations. This is due to the visual word mixture model, defined in Equation 5, where relative visual word frequencies are interpreted as visual word observation probabilities. Thus, it can be considered as a soft visual word observation model. For this reason the specificity of the pseudo-discrete HMM depends on the number of different visual words that are observed within an HMM state.

On the one hand, this has advantages for generalizing to unseen data. For example, in query-by-example word spotting where only a single exemplary instance of the query is given. Furthermore, no smoothing is required for parameter estimation, cf. Equation 4 and Equation 6.

On the other hand, this can lead to degenerated cases. When too many different visual words are observed within a certain state, the unimodal distribution looses its ability to discriminate different visual word configurations.

Another problem arises if non-discriminative, i.e., irrelevant, visual words keep reoccurring in different states and models. They tend to dominate the output probability due to the scalar product of HMM visual word probabilities and observed visual word frequencies in Equation 5.

*2) Exp. Dirichlet Compound Multinomial Model:* Particularly the last problem mentioned in the previous section (III-B1) has also been studied extensively for text analysis, cf. [15]. Non-discriminative terms occurring at higher frequencies have substantial influence in the models, e.g., in the multinomial distribution. Common heuristics for handling this are stop-word filtering or term weighting [15]. A more general approach, however, was found in the Dirichlet compound multinomial (DCM) distribution [15], [17]. This generative model consists of two stages where the parameters for the multinomial distribution are drawn from a Dirichlet distribution in the first stage and the BoW representation is finally drawn from the multinomial distribution in the second stage. This way the DCM has an additional degree of freedom that allows for modeling the concentration of visual words. Therefore, the model is not *"surprised"* [17] if words occur in bursts. Figure 1 shows a DCM probability mass function in a BoF probability simplex. The probability mass is mainly concentrated on the edges of the simplex, which indicates that high frequencies of few visual words are expected. Further details can be found in the caption of Figure 1.

Since the DCM is hard to estimate in large scale scenarios, Elkan [17] presented an exponential-family DCM approxi-

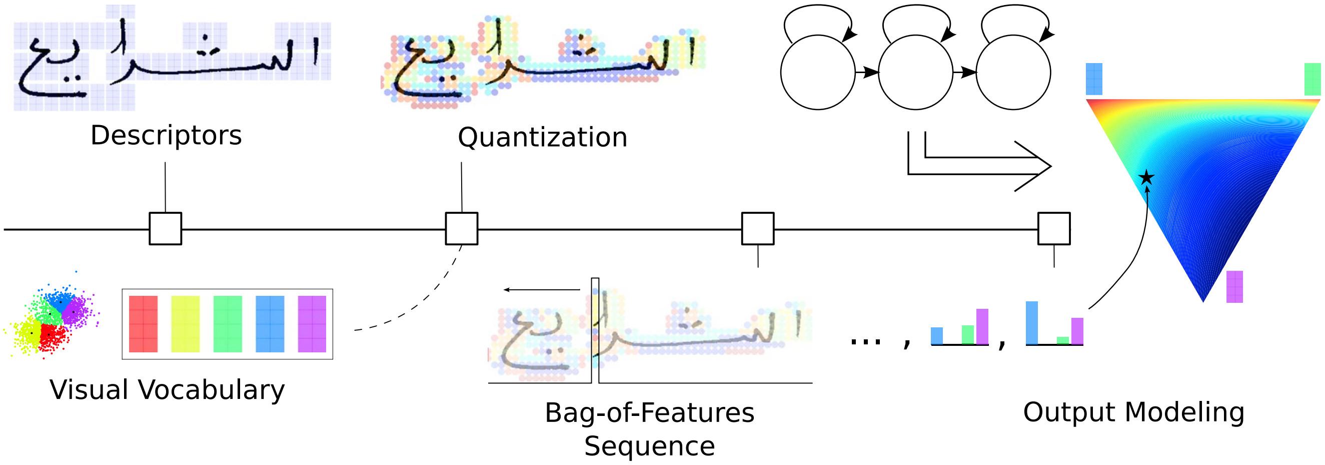

Figure 1. Word recognition with BoF-HMMs. From left to right, the feature extraction and modeling is shown examplarily. SIFT descriptors are computed in a word image. A few descriptors are visualized by blue patches. Afterwards, descriptors are quantized. For this purpose, descriptors are associated with their most similar visual word from a visual vocabulary. For visualizing the quantization result, colored descriptor center points are shown. The color encodes the index of the associated visual word. Visually similar pen strokes are covered with similar color patterns. In the next step, the BoF sequence is obtained by sliding a window over the word image in writing direction. This sequence is finally modeled within the HMM. The probabilistic model is shown examplarily as a function for three visual words over a probability simplex. In the simplex visualization each corner refers to a visual word and BoF can be represented as points within the simplex. The higher the (relative) frequency for a visual word in the BoF representation, the closer is the point to the respective corner. The BoF shown as a star, has high frequencies for the blue and pink visual words and low frequency for the green visual word. The probability mass function is indicated within the simplex with blue to red colors. Blue indicates low and red indicates high probability.

mation (EDCM) that is suitable for short text clustering. The model is also very suitable for our scenario. As in short text clustering, our feature representations are very high dimensional and extremely sparse.

Here, we integrate this model in a semi-continuous HMM where the observations $\mathbf{x}_t \in \mathbb{N}_0^V$ are absolute visual word frequencies. $\beta_{kv} \in \mathbb{R}_{\geq 0}$ are visual word specific concentration parameters for mixture component $k$ and $\Gamma$ denotes the gamma function:

$$b_j(\mathbf{x}_t) = \sum_{k=1}^{M} c_{jk} \left( n_t! \frac{\Gamma(s_k)}{\Gamma(s_k + n_t)} \prod_{v: x_{tv} \geq 1} \frac{\beta_{kv}}{x_{tv}} \right) \quad (7)$$

$$\text{with} \quad n_t = \sum_{v=1}^{V} x_{tv} \quad \text{and} \quad s_k = \sum_{v=1}^{V} \beta_{kv}.$$

It is interesting to note that the parameters $\beta_{kv}$ only have to be real positive numbers, while the parameters for the multinomial distribution (see Equation 3) are probabilities. This is the additional degree of freedom in the EDCM distribution, allowing for the modeling of visual word concentrations, i.e., visual word burstiness.

Further details regarding Equation 7 and EDCM mixture model parameter estimation can be found in [17]. Their model estimation can be directly integrated in Baum-Welch training by replacing the posterior probability of a mixture component with the posterior probability of selecting a mixture component at a given time in a given state.

In contrast to the unimodal pseudo-discrete HMM, the semi-continuous EDCM integration is better suited for representing many different writing styles in the same model. This can be an advantage for handwriting recognition.

## IV. Evaluation

We evaluate the two output models on a word recognition benchmark and on a segmentation-free query-by-example word spotting benchmark. This allows us to explore different properties of the models, as both tasks have very different requirements. We compare our results with the state-of-the-art on both benchmarks.

### A. Word Recognition

Word recognition is evaluated on the IfN/ENIT dataset of handwritten Tunisian town and village names [18]. It consists of subsets *a* - *f*, *s* and we use the common *training* - *validation* configuration *abc-d* and configurations *abcd-e*, *abcde-f* and *abcde-s* for *testing*. The dataset contains word images that exhibit great variations in writing style. Performance is measured by word error rate (WER). In our experiments we will investigate the robustness of the models with different feature configurations.

For recognition we extract sequences of BoF representations as described in Section III-A. A SIFT descriptor cell size of 13×13 pixels and a visual vocabulary of 4096 visual words performed best in the validation. The effect of different cell configurations (Desc. cells) and descriptor pruning (Desc. pruning) can be found in Table I. We estimate 178 context-dependent HMM character models with a Bakis topology. In case of BoF-HMMs with an EDCM output model, we estimate a mixture model with 1024 mixture

Table I
FEATURE ROBUSTNESS ON IFN/ENIT BENCHMARK (SET *abc* − *d*)

| BoF-HMM | Desc. cells (rows×columns) | Desc. pruning | WER (in %) |
|---|---|---|---|
| P.-Discrete | 4×4 | no | 26.5 |
| P.-Discrete | 4×4 | yes | 3.1 |
| P.-Discrete | 4×2 | no | 38.5 |
| P.-Discrete | 4×2 | yes | 2.4 |
| EDCM | 4×2 | no | 3.1 |
| EDCM | 4×2 | yes | 3.0 |

Table II
STATE-OF-THE-ART RESULTS ON THE IFN/ENIT BENCHMARK

| Method | *abc–d* | *abcd–e* | *abcde–f* | *abcde–s* |
|---|---|---|---|---|
| | ±0.5% | ±0.6% | ±0.6% | ±2.0% |
| Bernoulli-HMM [13] | 4.8 | 6.1 | 7.8 | 15.4 |
| Multi-Stage HMM [19] | 1.9 | 5.1 | 7.7 | 15.4 |
| BoF-HMM (p.-discrete) | 2.4 | 6.6 | 8.5 | 18.3 |
| BoF-HMM (EDCM) | 3.0 | 5.8 | 8.7 | 18.2 |

Table III
SIZE OF THE EDCM MIXTURE MODEL FOR WORD WORDSPOTTING

| Mixture components | mAP (in %) |
|---|---|
| 256 | 60.2 |
| 512 | 62.2 |
| 1024 | 62.4 |

Table IV
SEGMENTATION-FREE QUERY-BY-EXAMPLE WORD SPOTTING ON THE GEORGE WASHINGTON BENCHMARK

| Method | mAP (in %) |
|---|---|
| Exemplar SVM + Reranking [20] | 59.1% |
| BoF- Spatial Pyramid [4] | 61.4 % |
| BoF-HMM (p.-discrete) | 67.2% |
| BoF-HMM (EDCM) | 62.4% |

components according to the EM training scheme described in [17]. In Baum-Welch HMM training, only the state-dependent mixture weights are updated in order to avoid overfitting. The number of mixture components has experimentally been optimized on the validation set *abc-d*. WER converges with higher numbers of mixture components. This behavior is very similar to what we observed for EDCM model estimation in word spotting.

The robustness test is presented in Table I. It shows the effect of different feature configurations that mainly influence the specificity of the descriptors. As discussed in Section III-B1, the pseudo-discrete (p.-discrete) BoF-HMM degenerates if the same visual words keep reoccurring in many different states. We analyse the descriptor cell configuration and, most importantly, the descriptor pruning.

The standard setup for the SIFT descriptor is 4×4 cells [7]. However, for estimating character models it is advantageous if the horizontal context is limited. Results in Table I show that for the pseudo-discrete BoF-HMM vertical descriptor shapes only have a positive effect if unspecific descriptors are pruned. Otherwise, the model degenerates according to the specificity of the descriptors. Our most important result in Table I is that the EDCM model is completely robust in this regard. This is especially an advantage if heuristics for suppressing unspecific descriptors are hard to define.

Table II shows an overview of recent state-of-the-art results on the IfN/ENIT benchmark for validation / test sets *d, e, f* and *s*. The method presented by Ahmad et al. [19] achieves very good results by modeling the properties of Arabic script. Our method is, however, completely language and script independent and can be automatically adapted, cf. e.g., [4]. In comparison to the Bernoulli HMMs presented by Giménez et al. [13], our method achieves similar results on sets *d* and *e* while the Bernoulli HMMs generalize better in case of higher variability in writing style in sets *f* and *s*. This is most likely due to their less specific features.

### *B. Word Spotting*

Word spotting is evaluated on the George Washington benchmark, cf. [4], [6]. We consider a segmentation-free query-by-example scenario where the query is given as an exemplary occurrence of the query word. The query has to be retrieved without any prior information about word locations in the document images. The writing style in the handwritten documents is overall similar. Performance is measured by mean average precision (mAP), cf. [2]. In the segmentation-free evaluation scenario a detection is considered as relevant, if it overlaps with a corresponding annotation by more than 50%. In our experiments we show the effect of different numbers of mixture components of the EDCM model.

For representing document images, we extract standard SIFT descriptors with a cell size of 12×12 pixels and quantize them with respect to a visual vocabulary of 6144 visual words. Query word images can then be modeled with BoF-HMMs. For the pseudo-discrete output model, we directly estimate state specific visual word probabilities from the given sample. In case of the EDCM output model we initially estimate the mixture model with BoF sequences sampled from document images at typical line heights. At query time only the mixture weights are adapted. For detection we slide a patch through the document image and compute the probabilty that the underlying BoF sequence has been generated with the query model, cf. [6].

Table IV-B shows results for different EDCM mixture model sizes. It can be noted that the performance converges fast and that the model achieves good performance with a few hundred mixture components. This is positive as it shows that the model is very robust with respect to this parameter. Furthermore, this behavior was observed analogously for word recognition.

When comparing the performance of the EDCM model with the performance of the pseudo-discrete model in Table IV-B, the EDCM model is clearly outperformed. This is due to the characteristics of the historic document images that perfectly fit with the prerequisites of the pseudo-discrete model. The pseudo-discrete model has extremely high capabilities of generalizing to unseen data from a single sample. With respect to retrieval performance the state-of-the-art is clearly outperformed on the George Washington benchmark.

## V. Conclusion

In this paper we presented a study of different statistical models for BoF representations in handwriting recognition and word spotting. In order to model BoF sequences with HMMs in these scenarios, the special data characteristics have to be taken into account. Both, our pseudo-discrete BoF-HMM and the EDCM BoF-HMM are very well suited for high dimensional and sparse data. In our evaluation we showed that the EDCM model is very robust and performs well for word recognition and word spotting. The EDCM output model integration has to the best of our knowledge not been considered, before.

For the pseudo-discrete BoF-HMM we showed that descriptor pruning is required in the word recognition task. This is due to the data considered in this benchmark. Since the images are almost binary, the background is uniform. This results in visual words that are reoccurring in almost all states of the BoF-HMMs. In contrast, the document image background in the word spotting benchmark is textured and, therefore, represented with many different visual words. As the prerequisites for the pseudo-discrete BoF-HMM are met in this scenario, it reaches its full potential. We clearly outperform state-of-the-art retrieval accuracy for segmentation-free word spotting on the George Washington dataset.

## References

[1] S. O'Hara and B. A. Draper, "Introduction to the bag of features paradigm for image classification and retrieval," *Computing Research Repository*, vol. arXiv:1101.3354v1, 2011.

[2] J. Lladós, M. Rusiñol, A. Fornés, D. F. Mota, and A. Dutta, "On the influence of word representations for handwritten word spotting in historical documents," *IJPRAI*, vol. 26, no. 5, 2012.

[3] A. V. Ken Chatfield, Victor Lempitsky and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *Proc. British Machine Vision Conf.*, 2011, pp. 76.1–76.12.

[4] M. Rusiñol, D. Aldavert, R. Toledo, and J. Lladós, "Efficient segmentation-free keyword spotting in historical document collections," *Pattern Recognition*, vol. 48, no. 2, pp. 545 – 555, 2015.

[5] L. Rothacker, S. Vajda, and G. A. Fink, "Bag-of-features representations for offline handwriting recognition applied to Arabic script," in *Proc. of the Int. Conf. on Frontiers in Handwriting Recognition*, 2012.

[6] L. Rothacker, M. Rusiñol, J. Lladós, and G. A. Fink, "A two-stage approach to segmentation-free query-by-example word spotting," *manuscript cultures*, no. 7, pp. 47–57, 2014.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, 2004.

[8] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Int. Conf. on Computer Vision*, vol. 2, 2003.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[10] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. of the Int. Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 524–531 vol. 2.

[11] G. A. Fink, *Markov Models for Pattern Recognition, From Theory to Applications*, 2nd ed., ser. Advances in Computer Vision and Pattern Recognition. London: Springer, 2014.

[12] J. A. Rodrguez-Serrano and F. Perronnin, "A model-based sequence similarity with application to handwritten word spotting," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2108–2120, 2012.

[13] A. Giménez, I. Khoury, J. Andrés-Ferrer, and A. Juan, "Handwriting word recognition using windowed Bernoulli HMMs," *Pattern Recognition Letters*, vol. 35, pp. 149 – 156, 2014, frontiers in Handwriting Processing.

[14] P. Frasconi, G. Soda, and A. Vullo, "Hidden Markov models for text categorization in multi-page documents," *Journal of Intelligent Information Systems*, vol. 18, no. 2, pp. 195–217, 2002.

[15] R. E. Madsen, D. Kauchak, and C. Elkan, "Modeling word burstiness using the Dirichlet distribution," in *Proc. of the Int. Conf. on Machine Learning*, ser. ICML '05. New York, NY, USA: ACM, 2005, pp. 545–552.

[16] J. Yin and J. Wang, "A Dirichlet multinomial mixture model-based approach for short text clustering," in *Proc. of the Int. Conf. on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: ACM, 2014, pp. 233–242.

[17] C. Elkan, "Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution," in *Proc. of the Int. Conf. on Machine Learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 289–296.

[18] M. Pechwitz, S. S. Maddouri, V. Mrgner, N. Ellouze, and H. Amiri, "IFN/ENIT - database of handwritten Arabic words," in *Proc. of Colloque Int. Francophone sur l'crit et le Document*, 2002, pp. 129–136.

[19] I. Ahmad and G. A. Fink, "Multi-stage HMM based Arabic text recognition with rescoring," in *Proc. of the Int. Conf. on Document Analysis and Recognition*, 2015, pp. 751–755.

[20] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Segmentation-free word spotting with exemplar SVMs," *Pattern Recognition*, vol. 47, no. 12, pp. 3967 – 3978, 2014.