International Journal on Document Analysis and Recognition manuscript No.
(will be inserted by the editor)

Neural Models for Semantic Analysis of Handwritten Document Images

Oliver Tüselmann · Gernot A. Fink

Received: date / Accepted: date

Abstract Semantic analysis of handwritten document images offers a wide range of practical application scenarios. A sequential combination of handwritten text recognition (HTR) and a task-specific natural language processing system offers an intuitive solution in this domain. However, this HTR-based approach suffers from the problem of error propagation. An HTR-free model, which avoids explicit text recognition and solves the task end-to-end, tackles this problem, but often produces poor results. A possible reason for this is that it does not incorporate largely pre-trained semantic word embeddings, which turn out to be one of the most powerful advantages in the textual domain. In this work, we propose an HTR-based and an HTR-free model and compare them on a variety of segmentation-based handwritten document image benchmarks including semantic word spotting, named entity recognition, and question answering. Furthermore, we propose a cross-modal knowledge distillation approach to integrate semantic knowledge from textually pre-trained word embeddings into HTR-free models. In a series of experiments, we investigate optimization strategies for robust semantic word image representation. We show that the incorporation of semantic knowledge is beneficial for HTR-free approaches in achieving state-of-the-art results on a variety of benchmarks.

Keywords Document image analysis · Knowledge distillation · Document understanding

Oliver Tüselmann
Department of Computer Science, TU Dortmund University
E-mail: oliver.tueselmann@cs.tu-dortmund.de

Gernot A. Fink
Department of Computer Science, TU Dortmund University
E-mail: gernot.fink@tu-dortmund.de

1 Introduction

Natural language processing (NLP) enables machines to understand and process human language. Text analysis is a major area of research in this field, where semantic information is derived from purely textual input data. In recent years, text analysis approaches have made significant progress and are successfully deployed in a wide range of real-world applications [3, 17, 33].

Textual information is not necessarily encoded in a machine-readable format, but can also be represented as part of a handwritten document image. Semantic analysis of this input format is an interesting and challenging area of research due to the combination of visual and textual properties as well as the high variability of handwriting. Handwritten text recognition (HTR)-based approaches provide an intuitive realization in this domain consisting of a sequential combination of an HTR and a textual task-specific NLP system [10, 66]. In this process, the document image is first transformed into a machine-readable format using an HTR model, and then an NLP system is applied to the obtained text to solve the given task. The HTR and NLP models are trained independently, making it difficult or even impossible to correct recognition errors in the semantic model [14, 54].

To avoid the problem of error propagation, HTR-free models are well established [1, 47, 62]. These approaches are based on end-to-end neural architectures and avoid an explicit text recognition. The document image is transformed into a feature representation and a task-specific architecture is applied that uses vectorial rather than textual input to solve the semantic task. Even though the error propagation problem can be at least technically mitigated by HTR-free approaches, they have the fundamental disadvantage of

not being able to exploit important advances in NLP, such as pre-trained semantic word embeddings. Furthermore, specialized architectures based on vector instead of textual input are mandatory.

Both approaches have theoretical advantages and disadvantages when it comes to semantic analysis of handwritten document images. In this work, we present and compare an HTR-based as well as an HTR-free approach. Hereby, we evaluate both approaches on a variety of semantic handwritten document image benchmarks, including semantic word spotting, named entity recognition, and question answering. We identify the lack of pre-trained semantic word embeddings as a major problem of HTR-free approaches. Therefore, we propose a cross-modal knowledge distillation approach to efficiently transfer knowledge obtained in the textual domain into the visual one, without incorporating text recognition. A crucial issue in this integration process is the mapping of handwritten word images into a textually pre-trained semantic word embedding space using a convolutional neural network (CNN). We provide a detailed discussion of the challenges of performing such a mapping and present several approaches for optimizing it.

This work makes the following main contributions:

- We propose and compare an HTR-free and an HTR-based framework for semantic analysis of handwritten document images.
- We explore and evaluate optimization strategies for a robust semantic handwritten word image representation.
- We present a novel cross-modal knowledge distillation approach for HTR-free integration of pre-trained semantic knowledge from the textual to the visual domain.

This work summarizes previously published methods and results from [64–67] and extends them with new methods and further evaluations. In [65], we present an architecture for mapping word images into a textually pre-trained semantic word embedding space. We present an HTR-based approach for named entity recognition in [66] and an HTR-free approach for question answering in [67]. Recently, we published an evaluation of textual semantic word embedding approaches for a semantic representation of handwritten word images in [64].

2 Semantic Document Image Analysis

Semantic document image analysis is an interdisciplinary research area of computer vision (CV) and NLP. There are many relevant but specialized use cases in this area, ranging from automated grading of exams [48, 52] to understanding lecture notes [58]. In terms of real-world and generic applications, this work focuses on the tasks of semantic word spotting (see Section 2.1), named entity recognition (see Section 2.2), and question answering (see Section 2.3).

2.1 Semantic Word Spotting

Semantic word spotting realizes a semantic word image retrieval and can be seen as an extension of the traditional word spotting approach [70]. Given a query and a collection of documents, the goal of this task is to sort all word images from this collection according to their semantic similarity to the query. This allows users to search not only for word images with a particular transcription, but also for concepts which are latent or hidden inside a query [65]. There exists a variety of different query types with Query-by-Example (QbE) and Query-by-String (QbS) being the most prominent ones. In QbE applications, the query is a word image whereas in QbS it is a textual string representation. Furthermore, word spotting approaches can be divided into segmentation-free and segmentation-based [22]. In the former case, the entire document image is used without any segmentation at all, and in the latter case, an external segmentation at word or line level is required. An overview of semantic word spotting approaches from the literature is given below. For a detailed survey on traditional word spotting, see [22].

Early semantic word spotting approaches rely on ontology-based knowledge [29] and are thus limited to a small set of human labeled semantic relationships. To overcome this limitation, approaches based on textually pre-trained semantic word embedding models from the NLP domain have emerged [30, 65, 70]. These approaches rely on the common subspace representation strategy [5] and learn a mapping of word images into a textually pre-trained semantic space. Figure 1 illustrates this procedure. Given that the textual and visual data share a common representation, the similarity between a query and a word image is determined by their distance in embedding space. Early attempts for realizing this idea use a two-stage CNN-based approach [65, 70]. Thereby, the word images are converted into a feature representation and afterwards mapped into a semantic space using a fully connected neural

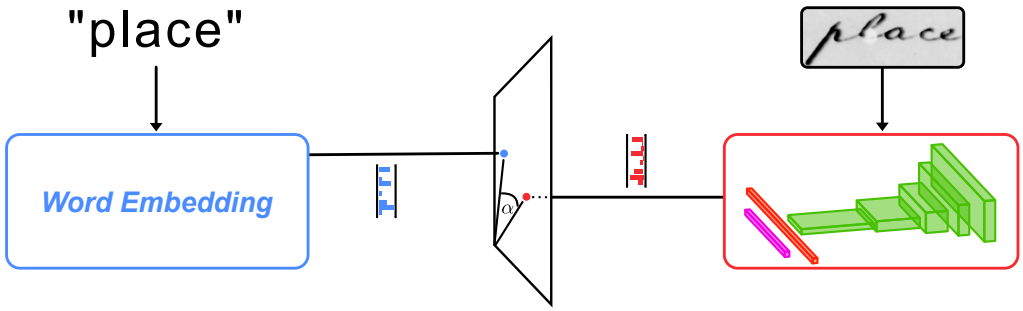

Figure 1: Visualization of the common subspace representation approach. A mapping of word images into a textually pre-trained semantic word embedding space is learned using a convolutional neural network. The similarity between two elements is determined by their distance (α) in this space.

network. The number of neurons in the last layer corresponds to the size of the semantic word embedding to be learned. End-to-end approaches are able to outperform two-stage architectures on semantic word image mapping [30, 63]. Due to the weak correlation between visual and semantic properties of words, the prediction of semantic word representations is much more difficult compared to classical word representations in the word spotting domain (e.g. PHOC [5]) [65]. Recently, the realization of a combined syntactic and semantic word image representation has been investigated [30, 63].

2.2 Named Entity Recognition

Named entity recognition (NER) is a sequence labeling task with a long tradition in NLP [74]. The goal of this task is to extract named entities (e.g. places, person, organizations) from an unstructured text. In the following, we will provide an overview of NER approaches on handwritten document images. A comprehensive review of NER in the text domain is given in [74] and for machine-printed document images in [15, 19].

A sequential NER approach offers an intuitive solution by first converting the given document into machine-readable text and then applying an NER model to the extracted text [21,40]. However, these approaches suffer from the error propagation problem [10, 18]. An alternative to the sequential approach is the use of integrated models that perform HTR and NER in a single step [12,18,43,59]. These approaches can be divided into segmentation-free [12, 46, 59] and segmentation-based [13, 43, 60]. Segmentation-free approaches are generally based on an encoder-decoder architecture. Here, the document image is first transformed into a two dimensional feature map. Based on this representation, separate task-specific decoders for HTR and NER are applied and jointly optimized [12]. Combining transcripts and named entity labels in a common output has proven to be a powerful alternative to separate prediction [13, 46, 59]. In general, bidirectional long short-term memory (BLSTM) [18] or Transformer [59] models are used for decoding and are trained on text recognition with special characters for named entities. Segmentation based approaches transform the given image of a text line into a two-dimensional feature representation using a CNN and encode the relationships between these features using a BLSTM model [13, 60]. Similar to the segmentation-free approaches, a nested prediction of the transcription and entity data is generated.

HTR-free models offer a promising solution for NER on handwritten document images. First approaches in this area focus on detecting named entities with handcrafted features [1]. A detection and classification of named entities based on word-level segmented document images can be achieved by a CNN, where the last layer corresponds to the number of entity classes [62]. This approach can be further improved by considering contextual information using an LSTM architecture [2, 48, 62]. In this approach, a CNN is used to extract features from the visual input and an LSTM for modeling the interactions between these features. Integrating additional information (e.g. part-of-speech tags) [47] or using an attention mechanism [2] leads to further improvements in this domain.

Recently, Transformer models for universal understanding of handwritten document images with self-supervised pre-training and autoregressive output generation were proposed [16, 27].

2.3 Question Answering

Search engines based on retrieval provide an efficient way to find information in large collections of data. A fundamental drawback of this approach is that users have to filter the results manually to find the answer

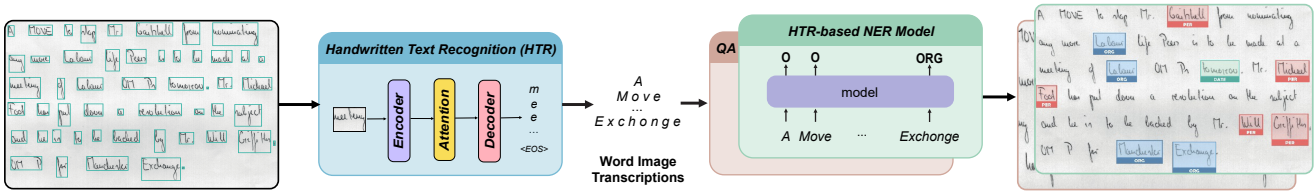

Figure 2: Overview of our proposed HTR-based framework for semantic analysis of handwritten document images. Each pre-segmented word image from the document is separately transcribed by a handwritten text recognizer. The resulting text is used as input to a task-specific NLP model, which fulfills the appropriate task.

to their query. To avoid this time-consuming and error-prone process, users demand that modern search engines provide a natural language answer to their query. This challenging task is known as question answering (QA) and has applications in a variety of disciplines [37, 72, 77].

In general, a distinction is made between extractive and generative QA models [8]. The extraction approach is also referred to as machine reading comprehension (MRC) and is based on the assumption that the correct answer to a question is available in the given context. Traditional MRC approaches are mainly based on handcrafted rules or statistical models [32, 76]. The use of neural approaches leads to major improvements over classical methods [8]. The common workflow of these neural approaches consists of an embedding, analysis, and prediction phase [8]. In the embedding phase, the context and question words are first transformed into vector representations. The interaction between these representations is modeled by a CNN, Transformer or recurrent neural network architectures [8]. In the prediction phase, a pseudo-probability distribution over the context words is determined to find the start and end positions of the answer. In addition to the classical pipeline, end-to-end architectures have been proposed in the literature [17, 33, 75]. These approaches encode the question and context words with a common Transformer model. A multi-layer perceptron (MLP) is applied to the context word representations to determine the start and end positions of the response [17]. Transformer models lead to state-of-the-art results on most MRC benchmarks and can be further improved by appropriate pre-training techniques [75].

There is an increasing interest in the QA task within the document image analysis community [31, 36, 37, 61]. The primary goal in this domain is to answer questions based on knowledge embedded in a given collection of document images. The multi-modal nature of document images makes this task particularly challenging. In addition to textual data, these images also contain structural and visual features that provide information that is relevant for answering questions. An important milestone in this area is the publication of the DocVQA dataset [37] and the Document Visual Question Answering Challenge [38]. Multi-modal approaches outperform sequential ones on nearly all QA benchmarks [6, 69, 73]. These approaches first extract visual, textual, and spatial information from a document image and transform them into feature representations. In most models, an external OCR approach is used to extract the textual and spatial information. The extraction of image features is based on well-established models from the CV domain, such as CNNs [6], U-nets [42], and region proposal networks [73]. These features serve as input to a Transformer encoder. This model is pre-trained on several self-supervised tasks with the goal of combining and matching the multi-modal features. For QA, an MLP is applied to the context word representations from the encoder in order to determine the start and end positions of the response [17].

Recently, Mathew et al. [36] published a first dataset for QA on handwritten document collections. Furthermore, they proposed an HTR-free QA approach, which outperforms HTR-based QA models on a historical handwritten dataset [36].

3 HTR-based Framework

A sequential combination of an HTR and a task-specific textual NLP model provides an intuitive solution for semantic analysis of handwritten document images. This method is referred to as HTR-based framework and is illustrated in Figure 2. All pre-segmented word images from the input image are separately converted into machine-readable text by an HTR model (see Section 3.1). The images are processed in the order they appear in the document. The generated text is used as input for an NLP model (see Section 3.2) to accomplish the intended semantic task. The HTR and NLP models are trained independently.

3.1 Handwritten Text Recognition

In this work, we use the attention-based sequence-to-sequence model proposed in [26] for HTR. This end-to-end model is based on the encoder-decoder paradigm and enables a transcription of handwritten word images. The encoder model extracts features from the word image using a combination of a CNN and a bidirectional gated recurrent unit (GRU). The decoder iteratively transcribes the given word image with an unidirectional GRU model and an attention mechanism.

Specifically, the handwritten word image I is transformed into a feature representation $X \in \mathbb{R}^{(N \times C \times D)}$ by a VGG network pre-trained on ImageNet [53]. For sequential processing, X is reshaped into a two dimensional matrix $X' \in \mathbb{R}^{(N \times K)}$. Motivated by the sequential nature of handwritten text, contextual representations $H \in \mathbb{R}^{(N \times J)}$ are computed from the columns of X' using a two layer bidirectional GRU model. The attention-based decoder uses the Bahdanau attention [7] to generate a context vector $c_t = \sum_{i=0}^{N-1} \alpha_{t,i} \cdot h_i \in \mathbb{R}^J$ at each decoding step $t \in 0, \ldots, T-1$. Here, $h_i \in H$ is a contextual feature representation of an image region and $\alpha_t \in \mathbb{R}^N$ is an attention vector. Based on the context vector, the decoder computes a pseudo-probability distribution with a unidirectional GRU model over the set of possible characters at each time step. The final output sequence $Y = (y_0, \ldots, y_{T-1})$ is generated by selecting the character with the highest pseudo-probability at each time step. For a detailed overview of this architecture, see [26].

3.2 HTR-based Semantic Models

A major advantage of the HTR-based approach is the possibility to use NLP models without adapting their architectures.

3.2.1 Semantic Word Spotting

We determine the semantic similarity between a query A and a word image B by the distance between their representations $\boldsymbol{a} \in \mathbb{R}^N$ and $\boldsymbol{b} \in \mathbb{R}^N$ in a textually pre-trained word embedding space from the NLP domain (e.g. FastText [9]). For this purpose, word images are transcribed by the HTR model and transformed into a vector representation using a semantic word embedding model. The similarity is calculated using the following formula:

$$dcos(A, B) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|} \tag{1}$$

A retrieval list is obtained by sorting all word images from the database in descending order according to their semantic similarity to the query.

3.2.2 Named Entity Recognition

The extracted machine-readable words from the document image are transformed into context-sensitive vector representations by a pre-trained RoBERTa model [33]. A two-layer BLSTM models the relationships between these representations. For each word image d_k a pseudo-probability distribution $\boldsymbol{y}_k$ over the possible entity classes C is generated. For this purpose, a combination of an MLP and a softmax function is applied to the outputs of the BLSTM. For word image d_k, the predicted entity class $\hat{y}_k$ is the corresponding class with the highest pseudo-probability in $\boldsymbol{y}_k$ (see Equation 2).

$$\hat{y}_k = \underset{i \in 1, \ldots, C}{\arg\max}\ \boldsymbol{y}_{k,i} \tag{2}$$

3.2.3 Question Answering

For efficiency, our HTR-based QA approach uses a combination of a retrieval and an answer extraction model. The goal of retrieval is to reduce the given document collection to a small number of documents that are relevant for answering a given question. The relevance of a document to the question is determined by the Term Frequency - Inverse Document Frequency (TF-IDF) score [34]. For the k most relevant documents, the answers to the question are computed separately. A QA architecture based on BERT [17] is used to extract the answer. This model is pre-trained on language modeling and is fine-tuned on the SQuAD dataset [44]. The answer is extracted from the given document by computing the start and end indices at token level, with all tokens between these two values representing the answer. In addition to the answer, a confidence score is obtained. The final answer of the system is the text passage of the k documents with the highest confidence.

4 HTR-free Framework

In this section, we present an HTR-free approach for semantic analysis of handwritten document images. Figure 3 illustrates this end-to-end framework, which avoids explicit text recognition and overcomes the problem of error propagation. The first step of this approach is to transform the pre-segmented word images from the input image into vector representations using a word image embedding model (see Section 4.1). The word images are processed independently in the order in which

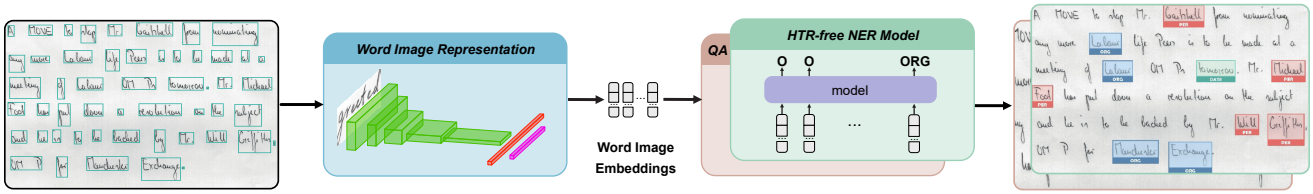

Figure 3: Overview of the proposed HTR-free approach for semantic analysis of handwritten document images. Each pre-segmented word image from the input image is transformed into a vector representation using a convolutional neural network. This sequence of representations is processed by a task-specific semantic HTR-free model.

they appear in the document. The obtained sequence of word image representations is then used as input for a task-specific semantic model (see Section 4.2) that solves the desired task.

4.1 Word Image Embedding

All word images of a given document have to be transformed into a vector representation of fixed dimensionality in order to be processed by a subsequent sequential model. In this work, we use a modified ResNet34 architecture [24] for word image representation. The global average pooling layer at the end of the network is replaced by a temporal pyramid pooling (TPP) layer [57]. The TPP layer produces a fixed-size representation based on the feature maps extracted by the CNN, while taking into account the sequential nature of handwriting. The final representation is obtained by applying a fully connected network to the output of the TPP layer. The dimensionality of the word image representation is fully customizable by the user. When incorporating knowledge from word representations, the dimensionality is determined by the representation to be predicted (e.g. FastText = 300).

4.2 HTR-free Semantic Models

Semantic analysis of document images without explicit text recognition requires appropriate architectures based on vectorial rather than textual input. The approaches presented in this work are mostly adapted NLP architectures.

4.2.1 Semantic Word Spotting

HTR-free semantic word spotting is accomplished by projecting word images and strings into a textually pretrained semantic word embedding space (see Figure 1). For machine-readable queries, the mapping is straightforward, whereas for word images, a mapping into this space must be realized. CNNs provide an excellent solution in the context of traditional word spotting approaches [28, 56] and are adopted for this task. Specifically, the CNN presented in Section 4.1 is used. The model is trained in a supervised manner by learning a mapping between word images and their associated semantic embeddings. These embeddings are provided based on the representations of their gold standard textual annotations in the pre-trained semantic space. Due to the embedding of text and image in the same vector space, the similarity between a query and a word image from the database can be determined by the distance of their representations in this space. In particular, the similarity score given in Equation 1 is used. A retrieval list is obtained that ranks all word images in the database by their similarity to the given query in descending order.

4.2.2 Named Entity Recognition

Our proposed HTR-free NER architecture is quite similar to the HTR-based one presented in Section 3.2. These approaches differ only in the word image representation part. The HTR-based model transcribes the word images of the given document and converts them into vector representations using a textual word embedding model. In contrast, the HTR-free model uses the CNN presented in Section 4.1 to convert each word image directly into a vector representation without transcribing it. End-to-end optimization is possible with this architecture.

4.2.3 Question Answering

For efficient HTR-free QA on large document collections, we propose a combination of a document retrieval and a QA model. The document retrieval model determines the relevance of each document in a given collection with respect to a given question. A comparison between the different modalities of word images and question words is achieved by the common subspace representation strategy. The relevance score $doc\_score(D, Q)$

of a document D for the question Q is calculated as follows:

$$doc\_score(D,Q) = \frac{1}{|Q|} \cdot \sum_{\boldsymbol{q} \in Q} \max_{\boldsymbol{w} \in D} [dcos(\boldsymbol{w}, \boldsymbol{q})] \quad (3)$$

Thereby, $\boldsymbol{w}$ and $\boldsymbol{q}$ are vector representations of the word images in D and the question words in Q, respectively. For each question word, the maximum similarity with respect to all word images from the document is computed. The final similarity score is the sum of these scores divided by the number of words in the question.

The k documents with the highest scores in the collection are processed individually by an answer extraction model. Our proposed QA approach is an adapted variant of the textual BIDAF architecture [50]. The textual inputs are replaced by vector representations of the input word images and the output format is changed from word to line level. Similar to the retrieval model, word images and question words are projected into a common subspace. Two separate BLSTMs are used to extract and model the relationships between the document and the question representations individually. Next, an attention-based representation is determined between the context and question word representations. The obtained representations are concatenated and serve as input to another two-layered BLSTM architecture that models the relationship between question and context words. In order to obtain the start and end rows of the answer, the BLSTM outputs are reduced to the number of lines in the document by summing the word representations according to their line membership in the document. A fully connected layer is applied to each of these line representations, which computes the start and end line indices. A confidence score of the prediction is obtained by the sum of the activations for the predicted start and end line indices. The final answer of the QA system is the prediction with maximum confidence. For a detailed overview of this architecture, see [67].

5 Cross-modal Knowledge Distillation

Powerful semantic word embedding models are available in the textual domain. However, such models are lacking for handwritten word images. The aim of this work is to transfer knowledge gained in the textual domain to handwritten word images without doing an explicit text recognition. Due to the HTR-free requirement, it is necessary to transfer knowledge across modalities. This process is known in the literature as cross-modal knowledge distillation [23] and is illustrated in Figure 4. In general, a teacher model is pre-trained on one input modality and its knowledge is transferred to a student model that uses a different modality. In our case, a pre-trained word embedding model from the NLP domain serves as the teacher and the CNN model proposed in Section 4.1 as the student. The input of the teacher is a machine-readable word and the input of the student is a handwritten word image. The student model is trained in a supervised manner to predict a semantic representation for a given word image that is similar to the one predicted by the teacher model using the textual annotation of that word image. This is achieved by using a loss function, called the distillation loss, which captures the difference between the prediction of the student model ($\hat{\boldsymbol{y}}$) and the teacher model ($\boldsymbol{y}$). The mean squared error (see Equation 4) is used as the loss function in this work. By minimizing this loss, the student model becomes better at making the same predictions as the teacher. During training, the gold-standard textual annotations of handwritten word images are available and used to create gold standard semantic word image representations based on the teacher model.

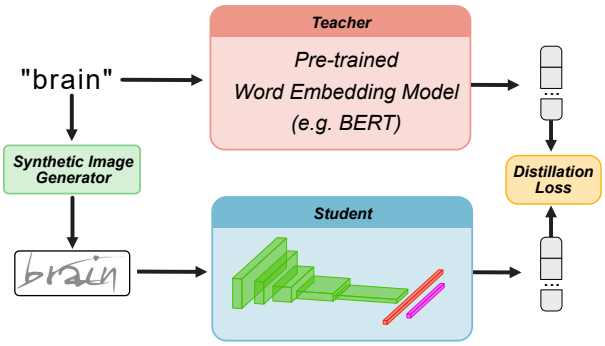

Figure 4: Overview of our knowledge distillation approach. Semantic knowledge is encoded by a textually pre-trained word embedding model. A semantic word image representation is learned by predicting a representation similar to the embedding of its annotation, which is obtained from the word embedding model. The distillation loss is a similarity measure of the representations obtained by the teacher and the student. The synthetic word image generator is only used in the annotation-free approach.

$$MSE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^2 \quad (4)$$

A crucial factor of this distillation approach is the choice of an appropriate textual semantic word embedding model. Therefore, we experiment with promising word embedding models from the NLP domain in Section 5.1. We will show that this distillation approach

has limitations for word images whose transcriptions did not occur in training. To address this issue, we present in Section 5.2 an annotation-free distillation strategy that uses synthetically generated word images to minimize the amount of out-of-training words. In addition, three strategies for integrating the semantic word image embedding model into the proposed HTR-free framework are presented in Section 5.3.

5.1 Semantic Word Embeddings

Semantic word embedding approaches use quite different approaches to encode semantic information. These methods can generally be divided into static and context-based [51]. A major limitation of static embeddings is that they are context-independent and thus ignore word polysemy. Context-based embeddings outperform static ones on almost all NLP benchmarks [20]. However, due to the independent representation of word images in our HTR-free approach, only static embeddings are suitable.

Word2Vec [39] is one of the first approaches for semantic word embedding. It uses a neural network to learn semantic representations for a given vocabulary based on the distribution hypothesis. The embeddings are generated independently of the word structure, making it impossible to predict embeddings for words that were not part of the vocabulary. To overcome this limitation, subword approaches such as FastText [9] and BytePair [25] have been proposed. These models split words into subwords and combine their embeddings into a single representation.

Context-based methods are used to encode word order information and tackle the problem of word polysemy. Even though context-based approaches are not directly suitable for our framework, static representations can be extracted that are able to outperform traditional static approaches on a variety of semantic benchmarks [20]. First context-sensitive models such as ELMO [41] and Flair [4] use LSTM architectures. A fundamental difference between these two approaches is that Flair processes the textual input purely character-based while ELMO uses a mixture of character and static word embeddings. Transformer-based encoders such as BERT [17] are trained on large text corpora in a semi-supervised manner. These encoders use a fixed size vocabulary and a tokenization approach to represent the input text. Furthermore, a positional embedding is used for word order encoding.

5.2 Annotation-free Knowledge Distillation

Knowledge distillation requires a sufficient amount of annotated training data [49]. This is particularly evident for our approach due to the lack of correspondence between semantic information and visual appearance of words. This missing link makes it generally impossible to predict the semantic information of an unknown word from its characters and their order. As a result, predicting semantic representations for word images whose annotations did not appear in the training set is difficult or even impossible [65]. Therefore, it is important to include as many words as possible in the training.

Unfortunately, the availability of large datasets with manually annotated word images are a major problem in the domain of handwritten document images. To address this problem, we propose an annotation-free knowledge distillation approach (see Figure 4) that relies on synthetically generated handwritten word images. This allows for efficient generation of large amounts of automatically labeled handwriting word images, and thus reduces the number of out-of-training words. The underlying assumption is that a sufficiently large set of synthesized word images can adequately cover the semantic space. Thus, only a few word image representations will be incorrectly predicted, which can be handled internally by an HTR-free model. Hereby, synthetic word images have already been successfully used in the handwriting domain to reduce or even eliminate the need for manually annotated data in various tasks (cf. e.g. [71]).

The main difference between the annotation-free approach and the traditional approach is at the input stage. The remaining workflow is identical to the traditional approach. Hereby, the annotation-free model requires just a machine-readable word as input, whereas the traditional model assumes a manually labeled word image. For the input word, a handwritten word image is synthesized using publicly available True Type fonts that resemble handwriting. Distortions and artifacts are randomly applied to the word image in order to generate realistic conditions. A vocabulary of input words is required. We use the most common English words as these are most likely to appear in future documents.

5.3 Integration Strategies

We investigate strategies for integrating distilled knowledge into our HTR-free approach. In the NLP domain, feature-based and fine-tuning approaches are well established for transfer learning [17]. The feature-based approach uses a pre-trained word embedding model to

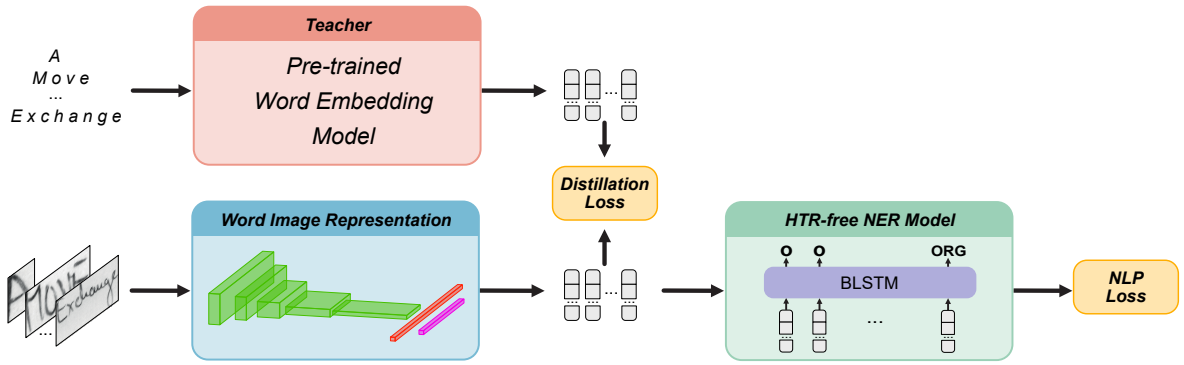

Figure 5: Overview of the multi-objective integration strategy. Besides training on the downstream task, the HTR-free approach is trained on knowledge distillation for the given handwritten word images. The NLP and distillation losses are jointly optimized.

transform input words into vector representations. These embeddings are used as input to a task-specific architecture. Thus, the word embedding model is not adapted during the training of the downstream task. In contrast, the fine-tuning approach adds a task-specific head to a given word embedding model and fine-tunes the entire model on the downstream task. In this way, the word embeddings can be adapted to the NLP task. For most NLP benchmarks, the fine-tuning approach can outperform the feature-based approach, but has substantial resource requirements [17].

We propose a feature-based and a fine-tuning approach for HTR-free semantic knowledge integration. By default, the weights of the HTR-free framework are randomly set. All proposed approaches initialize the word image embedding model with the weights of the distilled model. In the feature-based approach, the parameter of the embedding model are not adaptable during training. Thus, only the weights of the task-specific NLP model remain adjustable, and the pre-trained word image representations serve only as input. The fine-tuning approach is an end-to-end model, where the weights of the embedding model and the NLP model are tuneable. In both the feature-based and fine-tuning frameworks, the model is trained on the semantic downstream task in a fully supervised manner.

Due to the combination of a generally small number of annotated training examples for the downstream task and the large number of model parameters, overfitting to the training data can quickly occur. To alleviate this problem, we propose a multi-objective approach that extends the fine-tuning approach by optimizing the distillation loss in addition to the loss function for the NLP task. These two criteria are optimized by minimizing the sum of their losses. The multi-objective approach is illustrated in Figure 5.

6 Experiments

6.1 Datasets

We train and evaluate our proposed approaches on both synthetically generated as well as real handwritten word images from modern and historical documents. The datasets differ considerably in the size of the available training and test material, as well as in the number of writers. This allows to draw conclusion for a variety of real-world application scenarios. Bounding box information at word level is available for all datasets. In addition to their NER or QA specialization, the datasets are evaluated for HTR and semantic word spotting.

6.1.1 IAM Database

The IAM Database (IAM-DB) [35] is a major benchmark for HTR and word spotting. The documents contain modern English sentences and were written by a total of 657 different people. The database consists of 1539 scanned text pages containing a total of 13353 text lines and 115320 words. The official partitioning is writer-independent, such that each writer contributed to either the training, validation or test set. Manual named entity annotations and an optimized semantic partitioning are available [66]. These annotations are based on the established 18 categories from the OntoNotes Release 5.0 dataset. Two versions with 18 and 6 categories are provided. In the reduced version, the 18 labels have been summarized and severely underrepresented categories have been excluded.

6.1.2 Synth12K

The Synth12K dataset consists of synthetically rendered word images from True Type fonts that resemble

handwriting. This dataset is generated from a lexicon containing the 12000 most common English words. For each word, 50 training and 4 test images are generated. The training and test data use different fonts and are thus writer-independent. The font is randomly selected from over 300 publicly available fonts. Each word of the document is rendered onto a gray background. The font size and stroke intensity are randomly set.

6.1.3 HW-SQuAD

HW-SQuAD [36] is a QA dataset consisting of syntactically generated handwritten document images based on the textual SQuAD1.0 [44] dataset. The synthetic dataset consists of 20963 document pages containing a total of 84942 questions. The official partitioning splits the dataset into 17007 documents for training, 1889 for validation and 2067 for testing. The training, validation, and test sets contain 67887, 7578, and 9477 questions, respectively.

6.1.4 BenthamQA

BenthamQA [36] is a small historical handwritten QA dataset where questions and answers were created using crowdsourcing. The historic dataset contains 338 documents written by the English philosopher Jeremy Bentham and shows some considerable variations in writing styles. The dataset provides only a test set consisting of 200 question-answer pairs on 94 document images. The remaining 244 documents from the collection are used as distractors.

6.1.5 Synthetic Groningen Meaning Bank

The synthetic Groningen Meaning Bank (sGMB) dataset [12] consists of synthetically generated handwritten document pages obtained from the corpus of the Groningen Meaning Bank [11]. It contains unstructured English text mainly from a newspaper, whereby the words have been labeled with the following categories: *Geographical Entity, Organization, Person, Geopolitical Entity* and *Time indicator*. There is an official split containing 38048 training, 5150 validation and 18183 test word images.

6.1.6 George Washington

The George Washington (GW) dataset [45] has become the de-facto standard benchmark for word spotting. It consists of 20 pages of correspondences between George Washington and his associates dating from 1755. The documents were written by a single person in historical English. Manual named entity annotations are available as well as an optimized partitioning of the document images for semantic tasks [66]. The word images are manually labeled with the following categories: *Cardinal, Date, Location, Organization* and *Person*.

6.2 Evaluation Protocols

6.2.1 Semantic Word Spotting

Evaluation of semantic word spotting approaches requires both a semantic and a retrieval performance metric. Hereby, mean Average Precision (mAP) has been established as the de-facto standard metric for evaluating word spotting approaches. Specifically, the established protocol proposed in [5] is used. In the QbE setting, the first retrieved image is not included in the mAP calculation. For IAM-DB, only queries that are not part of the official stop word list are considered, but are kept as distractors during retrieval.

For evaluating the semantic quality of textual word embedding models, word analogy (WA) benchmarks have emerged. Here, manually predefined examples of semantic analogies are given which the model has to resolve. Formally, three words a, b, and c are provided in the WA task. The goal is to determine the fourth word d such that the following condition is satisfied: a is to b as c is to d. We follow the evaluation of [30] and use a collection of manually defined WA examples published in [39]. To adapt this metric to handwritten word images, we first compute embeddings for all word images in a given test set. For a given analogy, the textual semantic representations of words a, b, and c are determined. Then, the expected position $\hat{d}$ of the target word is computed:

$$\hat{d} = b - a + c \tag{5}$$

Based on this, the word image with the highest cosine similarity to $\hat{d}$ is determined. If the annotation of this word image matches the target word d, then the analogy is fulfilled. The percentage of correctly predicted analogies is used as the final measure of semantic evaluation. An analogy is discarded if the target word d does not have at least one word image with the same annotation in the test set.

6.2.2 Named Entity Recognition

The F1 score is the standard measure for evaluating NER models [74]. This score can be interpreted as a weighted average of precision and recall and is formally defined as shown in Equation 6. There are several definitions of this metric, with macro and micro F1 being the

most popular. In our experiments, we use the macro F1 score. This metric is computed independently for each entity class and finally averaged using the harmonic mean. Thus, all entity classes are considered equally, preventing the score from being dominated by a majority class.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{6}$$

6.2.3 Question Answering

The double inclusion score (DIS) [36] is an established metric for HTR-free QA systems. Inspired by the Intersection over Union (IoU) measure, this score determines the alignment of image regions from the predicted and gold standard answers. A formal definition of this metric is given in Equation 7. The small box (SB) contains all word images that are part of the gold standard answer. The large box (LB) contains all the word images from the lines that are part of the SB, as well as those from the lines above and below. The answer box (AB) contains the word images included in the image region predicted by a QA system. A prediction is considered correct if the DIS is greater than 0.8.

$$DIS = \frac{|AB \cap SB|}{|SB|} \cdot \frac{|AB \cap LB|}{|AB|} \tag{7}$$

6.3 Implementation Details

We do not make any changes to the hyperparameter and optimization strategy proposed in [26] for the HTR model. We only adjust the size of the input images, the maximum word length, and the alphabet for each dataset.

The word image embedding network is optimized with the mean squared error loss and the ADAM optimizer using a batch size of 64. The network is first pre-trained on the Synth12k dataset. A learning rate of 0.01 is used during pre-training and 0.001 during fine tuning. The images are scaled and padded to a fixed size of 128×384, while preserving the aspect ratio. Semantic representations are normalized to zero mean and unit variance.

In the proposed NER architecture, the BLSTM model uses a hidden layer size of 256 and a dropout of 0.5. For optimization, we use cross-entropy loss and the ADAM optimizer. The learning rate is initially set to 0.001 and is divided by two whenever the training loss does not decrease in a certain range within 10 epochs. We follow the label smoothing approach proposed by [62].

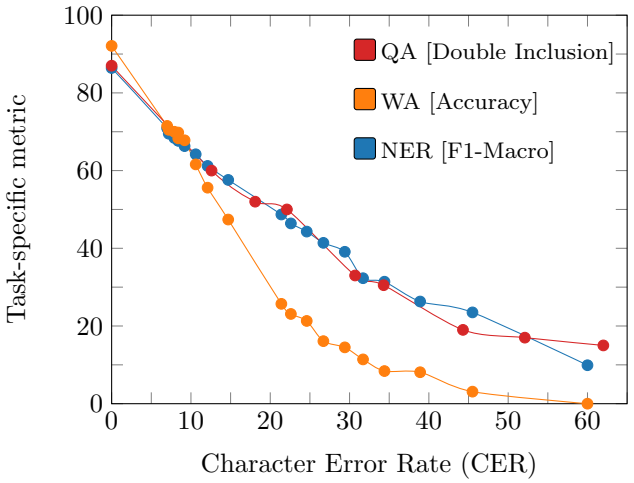

Figure 6: Impact of handwriting recognition errors on the performance of NLP approaches. The metric of each benchmark is given in brackets. The QA results are obtained on BenthamQA, the NER results on IAM-NER (6), and the WA on IAM-DB.

The BIDAF architecture is trained on HWSQuAD. We do not change the suggested parameters provided by [50]. The BLSTMs use a hidden layer size of 100 and a dropout probability of 0.2. ADADELTA is used to optimize the cross entropy-loss model with a learning rate of 0.5.

6.4 Effect of HTR Errors on NLP Models

In a first experiment, we investigate the impact of HTR errors on the performance of our proposed HTR-based framework. To achieve variable recognition rates, a variety of HTR models are trained with different percentages of training data ranging from 1% to 100%. These models are used to transcribe the word images from the test sets of the NLP benchmarks. The transcribed texts are subsequently used as input to a task-specific NLP model proposed in Section 3.2. The performance of these NLP models is further evaluated using perfect recognition results. Hereby, the annotations of word images are used as input to the NLP models.

Figure 6 illustrates the strong negative impact of HTR errors on state-of-the-art NLP approaches. With perfect recognition, high performances can be achieved on these benchmarks. However, the results for NER and word spotting degrade considerably even with a small number of text recognition errors (5% CER). There is a performance loss of about 15% on the NER benchmark and a loss of about 20% for semantic word spotting. In conclusion, the results indicate a strong negative in-

Table 1: Comparison of word embedding methods for named entity recognition on a variety of benchmarks. Results are reported in macro F1.

Embedding	IAM (6)	IAM (18)	GW	sGMB
Random	54.8	24.1	60.9	69.3
Syntactic	67.7	46.8	73.6	75.1
Semantic	87.5	63.5	89.6	80.2

fluence of HTR errors on the performance of textual semantic approaches.

6.5 Impact of Pre-trained Semantic Word Embeddings on Semantic Tasks

We consider the lack of pre-trained semantic knowledge to be a major limitation of HTR-free approaches. To test our assumption, we determine the relevance of pre-trained semantic word embeddings for NER. For this purpose, we replace the semantic word embedding used in our proposed NER model (see Section 3.2) with a randomly initialized word embedding and compare their performance on a variety of NER benchmarks. We evaluate the model under perfect text recognition by using the gold standard text annotations of the document image as input to the NER system. The randomly initialized word embedding method follows the idea proposed in [3]. Characters are converted into a randomly initialized but adaptable embedding of size 256. Words are formed by splitting the input string based on spaces. A BLSTM encodes a single word representation for each sequence of character representations. The final word embedding is obtained by concatenating the first hidden state of the backward model and the last hidden state of the forward model. We additionally evaluate a syntactic word embedding approach. Hereby, the pyramidal histogram of characters (PHOC) encoding [5] is used. This representation method encodes words with a small edit distance into a similar embedding, providing a kind of word recognition. For semantic word representation, we stick to the RoBERTa model.

Table 1 reveals large differences between the embedding methods used. Randomly initialized word embeddings can already correctly classify many named entities. Using syntactic word embeddings improves the performance considerably. This is probably due to the small amount of training data, where the given word structure increases the generalization ability and counteracts overfitting. Pre-trained semantic word embeddings are shown to be fundamental and lead to large performance improvements.

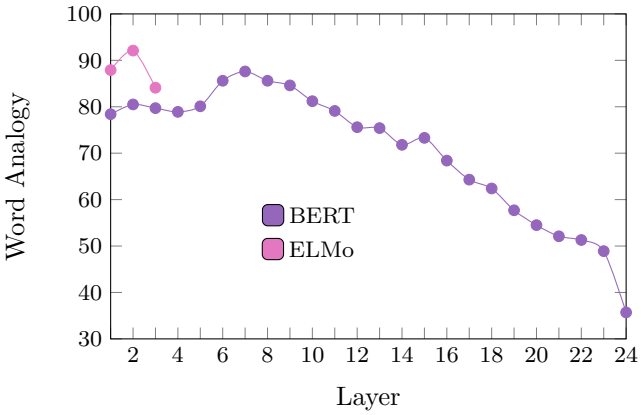

Figure 7: Performance of each layer from the BERT and ELMo models with respect to static word representation. The quality is determined based on a word analogy benchmark on the IAM-DB.

6.6 Knowledge Distillation

We evaluate our knowledge distillation approach for semantic word image representation. We first explore textually pre-trained word embedding models for acting as a teacher model. Then, we attempt to improve the robustness of distillation by using our annotation-free distillation approach. Finally, we evaluate our proposed integration strategies for using the distilled knowledge in our HTR-free framework.

6.6.1 Semantic Word Embeddings

Transforming machine-readable words into static vector representations is well defined for most approaches. However, there are several ways to extract static word representations from BERT and ELMo models. In order to select an appropriate static word representation, Figure 7 visualizes the semantic quality of each layer in the respective models. The WA score on the IAM benchmark is used as semantic quality measure. For the BERT model, the performances of the individual layer vary considerably, with the first layers of the model realizing a powerful static word representation. In contrast, the last layers show low performance on the static benchmark and seem to encode mainly contextual information. The ELMo model has only three output layers, whose results on the benchmark vary only slightly.

Figure 8 (a) illustrates WA scores on the gold standard text annotations of IAM-DB for a variety of word representation methods. The results highlight the advantages of semantic over syntactic word representations in encoding semantic knowledge. Semantic embeddings are able to resolve more than 90% of the

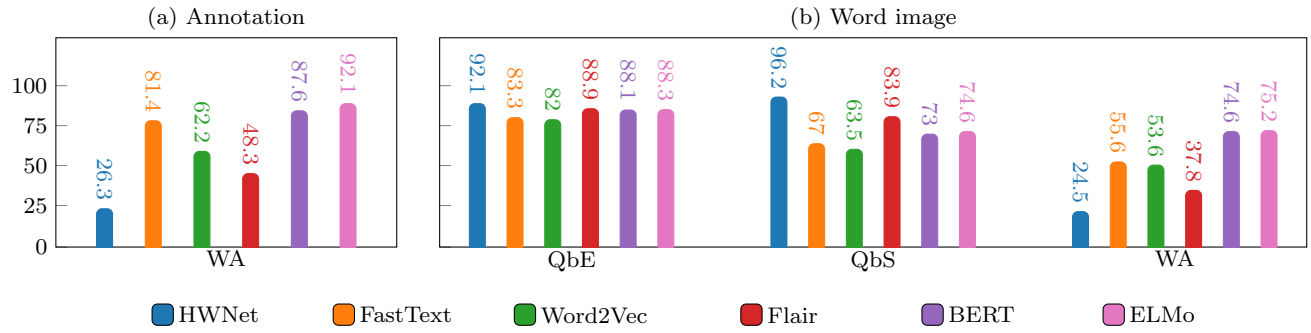

Figure 8: Textual word representation methods in terms of their ability to represent semantic information for handwritten word images. An upper bound on the semantic quality based on perfect text recognition is given in (a). The quality of the representations based on handwritten word images is given in (b), where QbE, QbS, and WA are used as quality measures. Results are reported on IAM-DB.

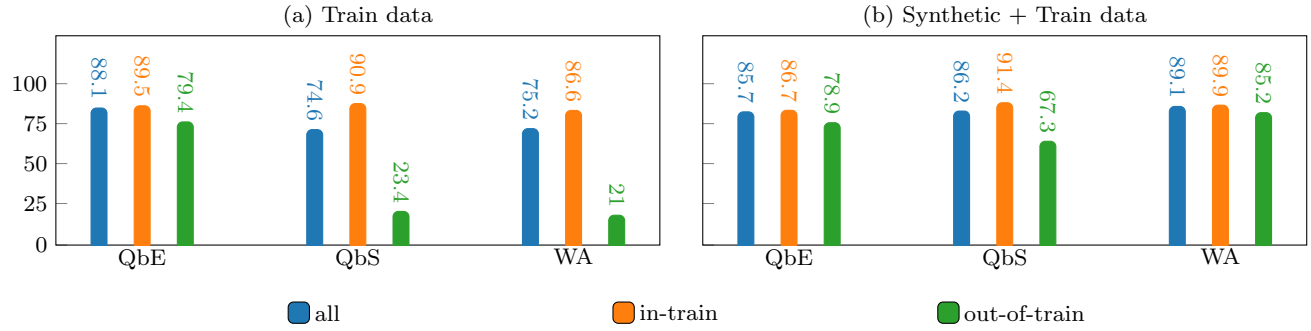

Figure 9: Quality of predicting semantic word image representations based on their occurrence in the training set. In-train represents all queries or analogies that occurred as annotations in the training, and out-of-train represents all other cases.

given word analogies, while syntactic embeddings are able to resolve only about 25% of these analogies correctly. Static embeddings extracted from the context-based BERT and ELMo approaches considerably improve the results compared to classical methods such as FastText and Word2Vec. The purely character-based Flair embedding leads to a comparatively low performance.

Besides the semantic quality of word embeddings, their prediction based on word images is relevant for our approach. In general, the experimental results in Figure 8 (b) show that the choice of an appropriate textual word embedding method has a fundamental impact on the quality of semantic word image representation. Although there is little variation between the embedding methods in terms of QbE values, the advantages and disadvantages of these methods are evident in terms of QbS and WA scores. The syntactic PHOC representation can be predicted well based on word images, but encodes little to no semantic information. Although a purely character-based Flair embedding improves the semantic quality, it also leads to a reduced performance in terms of prediction. As expected, Word2Vec achieves the lowest performance in the QbS benchmark due to the missing link between word shape and representation. FastText can improve performance over Word2Vec by taking n-gram information into account. BERT and ELMo models provide the best trade-off between semantic quality and prediction based on word images. Even though the BERT embedding shows only minor differences compared to the ELMo representation on IAM-DB, ELMo provides the best ratio between WA and QbS scores on all benchmarks tested. Therefore, the ELMo embedding is used as semantic word image representation method in the remaining experiments.

6.6.2 Annotation-free Knowledge Distillation

The quality of predicting semantic word image representations lags far behind syntactic ones. We analyze the semantic prediction in Figure 9 (a). The results illustrate that the quality of predicting a semantic word image embedding strongly depends on whether the annotation of a given word image was part of the training or not. First, we evaluate the coverage of test words in IAM-DB. Using only the training data, 54.7% of the

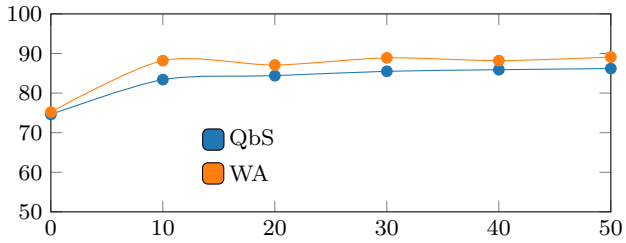

Figure 10: Influence of synthesized word images on predicting ELMo representations. The performance of the predictions is evaluated using WA and QbS benchmarks. A value of 0 on the x-axis represents the set of words from the training data, and any value $x > 0$ represents the inclusion of the $x \cdot 1000$ most frequent English words.

test words appear in the training set. By including synthetic word images of the most frequently used English words, the coverage of the test words can be increased to more than 90%. Around 80% of test words is already covered by including 10.000 words.

Figure 10 highlights the positive influence of integrating synthetically generated word images during training. There is a high correlation between the coverage of the test words and the prediction performance of the semantic embeddings. The largest performance increase is obtained when the 10.000 most frequent words are used. Unfortunately, the performance does not improve considerably with an increasing number of synthesized words. The improvements on the QbS and WA benchmarks can be attributed to the enhanced prediction of words that did not occur in training. This is shown in Figure 9 (b).

6.6.3 Evaluation of Integration Strategies

In order to integrate the distilled semantic knowledge into the HTR-free framework, we evaluate and compare the methods presented in Section 5.3. Figure 11 shows results for the three integration strategies on a variety of NER and QA benchmarks. Additionally, where possible, a random initialization of the word embedding model is evaluated as a baseline approach.

Compared to random initialization, the three presented integration approaches lead to considerably improved results on NER benchmarks. Performances of the randomly initialized models differ across the benchmarks. For the three integration approaches, the results are almost identical, with the multi-objective approach performing best and the fine-tuning method worst.

Due to the machine-readable format of the question, a QA model with randomly initialized word embeddings is not appropriate. For both evaluated QA benchmarks,

Table 2: Performance of our HTR models on the evaluation benchmark datasets. Results are given in CER and WER.

Dataset	CER	WER
IAM (WS)	5.5	14.3
IAM (SEM)	7.0	19.9
GW	3.1	8.0
sGMB	2.7	9.1
BenthamQA	19.1	43.8
HWSQuAD	0.5	1.7

the fine-tuning approach achieves considerably worse results than the feature-based approach. This is probably due to the more complex architecture and problem definition with respect to NER. Given the combination of complexity and huge amount of adjustable parameters of the model, overfitting to the training data could be the reason for poor generalization. In addition, the multi-objective method is not suitable for BenthamQA. This is explained by the unrepresentative nature of the synthetically generated word images between training dataset to the historical word images from the test set. This causes a considerable degradation in the prediction of semantic word representations on the test data.

Overall, the multi-objective approach provides the best results for all evaluated NER benchmarks, while the feature-based approach yields the best results for the QA benchmarks. Considering the only slightly worse scores for the feature-based approach on the NER benchmarks and the considerably lower resource requirements for training compared to the end-to-end approaches, we recommend the use of the feature-based approach in practice.

6.7 Task-specific Evaluation

After setting the appropriate hyperparameters for our HTR-free approach, we evaluate our proposed frameworks and compare them with approaches from the literature on a variety of benchmarks. These benchmarks include semantic word spotting, named entity recognition, and question answering. Hereby, our HTR-free approach uses ELMo as semantic word image representation, annotation-free knowledge distillation as pretraining and the feature-based integration strategy. Table 2 shows the text recognition performance of the HTR models on the benchmark datasets.

6.7.1 Semantic Word Spotting

There are currently only a few publications in the literature related to semantic word spotting. These methods

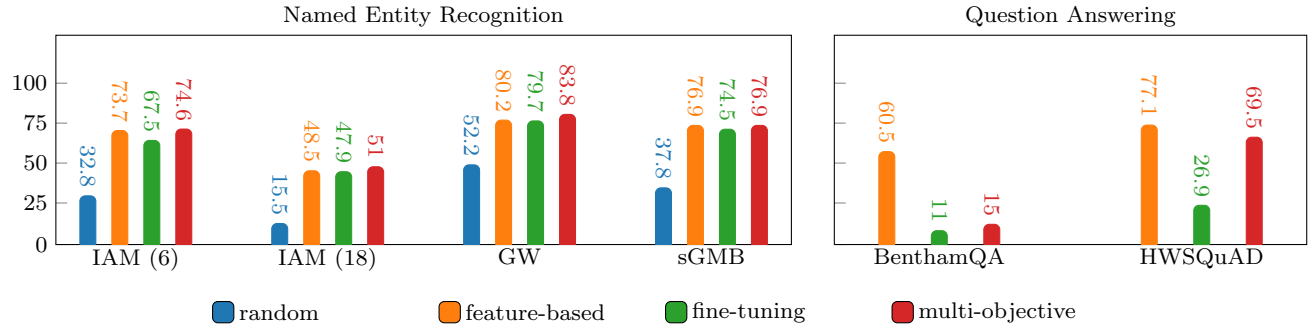

Figure 11: Performance of the proposed integration strategies on a variety of named entity recognition and question answering benchmarks.

Table 3: Comparison of our proposed HTR-free and HTR-based approaches on QbE and QbS semantic word spotting. Results are given in mAP. Furthermore, the semantic quality is evaluated by a WA task.

	GW			*IAM-DB*		
Approach	QbE	QbS	WA	QbE	QbS	WA
PHOCResNet [55]	97.8	98.0	—	85.5	94.1	—
HWNetv3 [28]	99.5	99.8	—	93.2	97.5	—
Triplet-CNN [70]	96.9	69.8	—	81.6	75.7	—
Sem-MSE [30]	97.6	94.4	—	84.6	69.7	63.2
Sem-Rank [30]	97.8	93.7	—	83.3	71.3	65.6
Combined [30]	**99.4**	98.8	—	**90.6**	**94.3**	61.5
HTR-based	97.0	96.6	**100.0**	77.3	84.0	73.7
HTR-free	98.1	**98.9**	**100.0**	85.7	86.2	**89.1**

Table 4: Additional semantic word spotting evaluation on the HWSQUAD, BenthamQA and sGMB datasets. QbE and QbS results are reported in mAP and WA in accuracy.

	HTR-free			*HTR-based*		
Dataset	QbE	QbS	WA	QbE	QbS	WA
HWSQuAD	**99.5**	96.9	**84.2**	98.9	**98.9**	83.8
BenthamQA	**72.8**	**74.6**	**49.7**	49.1	56.5	35.1
sGMB	**94.2**	**91.5**	82.3	89.1	91.3	**84.3**

have only been evaluated for IAM-DB and GW. Table 3 provides results obtained by state-of-the-art semantic and syntactic word spotting models, as well as our proposed HTR-free and HTR-based approaches. Especially on IAM-DB, the difference between the HTR-free and the HTR-based model becomes evident. The performance of the HTR-based approach for QbE retrieval and WA task decreases. Compared to approaches in the literature, our proposed HTR-free method can considerably improve the WA performance, especially on IAM-DB. Even though a combination of semantic and syntactic representations as proposed in [30] leads to better QbS results, the approach shows lower semantic quality and generally provides a worse trade-off between QbS and WA metrics.

Table 5: Comparison of our proposed frameworks with NER approaches from the literature on several NER benchmarks. Results are given in macro-F1. In addition, a baseline of the HTR-based approach working on gold standard text annotations of the input word images is evaluated.

Approach	IAM (6)	IAM (18)	GW	sGMB
Toledo et al. [62]	37.4	18.0	45.3	38.8
Carbonell et al. [12]	—	—	—	53.5
Rowtula et al. [47]	54.6	30.3	66.6	60.1
Dessurt [16]	71.1	48.5	—	—
Line-level CN [68]	57.2	46.5	68.8	—
HTR-based	**76.4**	**53.6**	81.3	75.8
HTR-free	74.6	51.0	**83.8**	**76.9**
HTR-based (Ann.)	87.5	63.5	89.6	80.2

We further evaluate our approaches on sGMB, HWSQuAD and BenthamQA in Table 4. For HWSQuAD and sGMB, there is no clear advantage for any of our proposed approaches. This is probably due to the low CER scores on both datasets. However, on the challenging BenthamQA dataset, the HTR-free approach provides considerable performance improvements on all metrics. This is due to the large number of text recognition errors produced by the HTR model, resulting in many incorrectly predicted semantic word embeddings.

6.7.2 Named Entity Recognition

Table 5 compares our proposed approaches to methods from the literature on a variety of NER bechmarks. A comparison of our HTR-free and HTR-based framework reveals that both approaches achieve similiar results on all benchmarks. The HTR-based approach per-

Table 6: QA results obtained on HWSQuAD and BenthamQA datasets. Results are given independently for retrieval (R), answer extraction (E), and the entire pipeline (C). In addition, a baseline of the HTR-based approach working on gold standard text annotations of the input word images is evaluated.

	HWSQuAD			*BenthamQA*		
Approach	R	E	C	R	E	C
Mathew et al. (HTR-free) [36]	46.5	—	15.9	55.5	—	17.5
Mathew et al. (HTR-based) [36]	86.1	—	59.3	32.0	—	2.5
BIDAF-Line [67]	86.2	68.1	45.0	92.5	50.5	37.5
HTR-based	**89.7**	**95.1**	**73.5**	85.5	50.0	41.5
HTR-free	87.0	77.1	53.3	**94.5**	**60.5**	**51.0**
HTR-based (Ann.)	90.0	96.0	74.4	98.5	87.0	80.0

Table 7: Comparison of direct and sequential word image embedding approaches on several NER benchmarks with the macro-F1 score. All models use the same HTR-free NER architecture and embed the input word images into ELMo representations. HTR-free directly predicts the ELMo representation of a given word image using a CNN. HTR-free (HTR) first transcribes the word image with an HTR model and uses the extracted text to obtain its ELMo embedding. HTR-free (Ann.) acts as a baseline approach that uses the text annotation of the word image for ELMo embedding.

Approach	IAM (6)	IAM (18)	GW	sGMB
HTR-free (Ann.)	85.7	63.4	84.5	81.1
HTR-free (HTR)	72.1	46.9	81.9	75.2
HTR-free	74.6	51.0	83.8	76.9

forms slightly better on both versions of IAM-DB and the HTR-free approach on GW and sGMB. Compared to an oracle approach with perfect text recognition as input, the performance of our methods shows a huge decrease, especially on the IAM-DB benchmarks. When compared to approaches from the literature, our methods show considerable improvements. Dessurt [16] still performs well, probably due to the implicit encoding of semantic knowledge extracted by the self-supervised pre-training. The models proposed in [47, 62] have a very similar NER architecture, but they do not integrate external semantic knowledge. Instead, they train the model end-to-end with randomly initialized weights. This again highlights the benefits of our cross-modal knowledge distillation approach.

6.7.3 Question Answering

Table 6 shows a comparison of our proposed QA approaches and methods from the literature on HWSQuAD and BenthamQA. The results highlight the vulnerability of models with explicit text recognition in presence of low HTR performance. Here, HTR-based approaches can only achieve low performance on BenthamQA, despite the use of a state-of-the-art QA model. However, at a reasonable recognition rate, the advantages of the textual QA model become apparent and lead to almost perfect results on HWSQuAD. The oracle approach performs well on the retrieval and extraction tasks independently, but the overall pipeline performance is rather low at 75% on HWSQuAD. This is likely due to the primary intent of the dataset, which was to extract the response from a single document rather than a collection of documents. This is supported by the performance on BenthamQA, which was designed specifically for document collections and has a more realistic relationship between performance on each subtask and their combination. In [67], we use the same modified BIDAF architecture as in our proposed HTR-free approach, but instead of using a semantic representation, we use PHOC for word image representation. There are clear advantages in integrating semantic knowledge, especially for the answer extraction task in both benchmarks. Compared to the HTR-free approach proposed in [36], large performance gains can be achieved by our HTR-free method. In particular, the HTR-free retrieval of relevant document images shows high robustness. The performance of our HTR-based approach is considerably higher compared to the method proposed in [36]. This is likely due to the improved robustness of our HTR model, which is achieved by integrating samples of IAM and GW during training.

7 Discussion

7.1 Direct vs. Sequential Word Image Embedding

In the following analysis, we demonstrate the advantages of directly predicting a semantic embedding for a word image instead of sequentially recognizing and embedding it. Both approaches have the same conditions and use the same NER architecture as described in Section 4.2. Furthermore, they rely on a static version of the ELMo embedding for semantic word image representation. As shown in Table 7, a direct embedding performs better on all NER benchmarks tested. As a baseline approach, we evaluate a system under perfect prediction of semantic ELMo embeddings. For this purpose, the gold standard text annotations of the

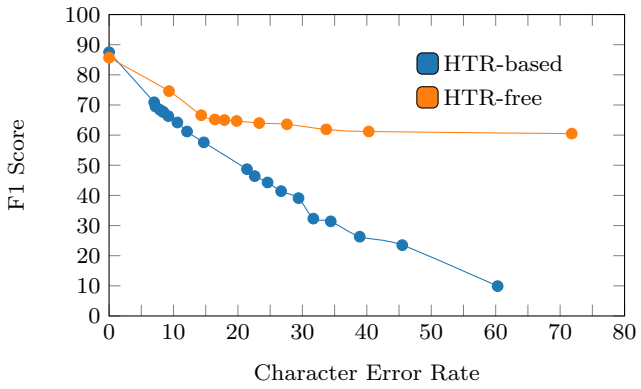

Figure 12: Assessing the robustness of the HTR-free and HTR-based approaches. Results are shown for both approaches on the IAM (6) NER benchmark as a function of CER. HTR and word embedding models are pre-trained on different amounts of data from the IAM-DB, resulting in varying CERs.

word images are used to generate semantic word embeddings. Finally, these embeddings serve as input for the HTR-free NER model. Even if the performance of the oracle is close to the performance of the direct and sequential approaches on GW and sGMB, a large difference is observed on IAM-DB. Thereby, the text recognition error rate for GW and sGMB is quite low with about 3% CER compared to IAM-DB with 7% CER. This again shows the strong influence of HTR errors on the performance of NER tasks.

7.2 Robustness of HTR-free Approach

High robustness to handwriting variability is an established expectation when using HTR-free models. To test this assumption, we evaluate and compare the robustness of our proposed frameworks. In this experiment, we obtain the performance of both approaches on an NER benchmark as a function of CER. We train HTR and word image embedding models on different subsets of the training data to achieve different recognition performances. For the HTR-free approach, a lexicon-based recognition is performed to compute CER. The lexicon consists of all training, validation and test words of the given benchmark dataset. Based on these HTR and embedding models, transcriptions and semantic word image embeddings are calculated respectively. For the HTR-based approach, the transcriptions are used as input to the textual NER model, and the word embeddings are used as input to the HTR-free NER model.

Figure 12 shows their results on the IAM (6) NER benchmark. The results demonstrate the fundamental disadvantage of sequential approaches, as HTR errors propagate and lead to poor performance, especially at high CER. The HTR-free model, on the other hand, uses the multi-objective approach and is thus able to achieve high performance for NER even when the initial word embedding model has poor performance. While the performance of the HTR-free approach drops considerably in the beginning, the multi-objective technique is able to correct or at least cope with the incorrectly predicted word image embeddings for high CERs. This is probably due to its end-to-end architecture and optimization strategy.

Table 8: Ablation study of our optimization strategies for semantic word image embedding on NER benchmarks. Results are given in macro F1.

Approach	IAM (6)	IAM (18)	GW	sGMB
Random	32.8	15.5	52.2	37.8
PHOC	60.3	36.5	72.8	71.2
FastText	53.6	32.3	68.2	66.9
ELMo	63.9	40.2	72.1	70.2
ELMo (+Synthetic)	74.6	51.0	83.8	76.9

7.3 Ablation Study

Considering the ambiguous correlation of semantic word image embedding and task-specific results, the effects of the most relevant optimization steps are evaluated in an ablation study. Table 8 shows the effects of our proposed optimization steps on NER benchmarks. Even if the performance on these benchmarks varies considerably, a general trend of optimization effects is observed. Thereby, an appropriate word image representation can almost double the performance on all benchmarks compared to the standard random initialization technique. The results also show that the predictive quality is at least as important as the encoding of semantic information. Hereby, the syntactic PHOC representation outperforms the semantic FastText embedding on all benchmarks, even though PHOC encodes almost no semantic information. The ELMo representation has both high semantic quality and good predictive ability on handwritten word images. Especially when pre-trained on synthetic data, the ELMo representation achieves the best results on all benchmarks. Overall, there is a strong correlation between the semantic word image embedding and task-specific metrics. However, the performance of NER approaches cannot be determined by a single metric, but requires a joint consideration of QbS and WA scores.

8 Conclusions

In this work, we present an HTR-free and an HTR-based framework for semantic analysis of handwritten document images. We evaluate the proposed approaches on a variety of semantic benchmarks involving semantic word spotting, named entity recognition, and question answering. The HTR-based approach suffers from error propagation, but outperforms HTR-free approaches on most benchmarks. We identify the lack of pre-trained semantic word embeddings as a major problem of HTR-free approaches. We propose a cross-modal knowledge distillation approach to efficiently integrate semantic knowledge from textually pre-trained word embedding models into the HTR-free framework, while avoiding explicit text recognition. This leads to a considerable performance improvement over classical HTR-free models and reduces the gap to state-of-the-art approaches. A crucial issue in this integration process is the mapping of handwritten word images into a textually pre-trained semantic word embedding space. While our proposed optimization methods have made great progress in this mapping process, the task remains challenging.

The approaches proposed in this work require that the input document images are segmented at word level. In future work, we attempt to address this limitation by extending our approaches to handle input images at line or paragraph level. Furthermore, we will investigate the replacement of static word embeddings by contextualized ones in our HTR-free framework.

References

1. Adak, C., Chaudhuri, B.B., Blumenstein, M.: Named entity recognition from unstructured handwritten document images. In: Int. Workshop on Document Analysis Systems, pp. 375–380 (2016)
2. Adak, C., Chaudhuri, B.B., Lin, C., Blumenstein, M.: Detecting named entities in unstructured Bengali manuscript images. In: Int. Conf. on Document Analysis and Recognition, pp. 196–201 (2019)
3. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: Annual Conf. of the North American Chapter of the Association for Computational Linguistics, pp. 54–59 (2019)
4. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Int. Conf. on Computational Linguistics, pp. 1638–1649 (2018)
5. 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)
6. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: DocFormer: End-to-end transformer for document understanding. In: Int. Conf. on Computer Vision, pp. 973–983 (2021)
7. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Int. Conf. on Learning Representations (2015)
8. Baradaran, R., Ghiasi, R., Amirkhani, H.: A survey on machine reading comprehension systems. Natural Language Engineering **28**(6), 683–732 (2022)
9. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
10. Boros, E., Romero, V., Maarand, M., Zenklová, K., Krecková, J., Vidal, E., Stutzmann, D., Kermorvant, C.: A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. In: Int. Conf. on Frontiers in Handwriting Recognition, pp. 79–84 (2020)
11. Bos, J., Basile, V., Evang, K., Venhuizen, N., Bjerva, J.: The groningen meaning bank. In: Joint Symposium on Semantic Processing, pp. 463–496 (2017)
12. 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)
13. Carbonell, M., Villegas, M., Fornés, A., Lladós, J.: Joint recognition of handwritten text and named entities with a neural end-to-end model. In: Int. Workshop on Document Analysis Systems, pp. 399–404 (2018)
14. Chiron, G., Doucet, A., Coustaty, M., Visani, M., Moreux, J.: Impact of OCR errors on the use of digital libraries: Towards a better access to information. In: Joint Conf. on Digital Libraries, pp. 249–252 (2017)
15. Cui, L., Xu, Y., Lv, T., Wei, F.: Document AI: Benchmarks, models and applications. CoRR **abs/2111.08609** (2021)
16. Davis, B.L., Morse, B.S., Price, B.L., Tensmeyer, C., Wigington, C., Morariu, V.I.: End-to-end document recognition and understanding with Dessurt. In: European Conf. on Computer Vision, pp. 280–296 (2022)
17. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Annual Conf. of the North American Chapter of the Association for Computational Linguistics, pp. 4171–4186 (2019)
18. Dhiaf, M., Jemni, S.K., Kessentini, Y.: DocNER: A deep learning system for named entity recognition in handwritten document images. In: Int. Conf. on Neural Information Processing, pp. 239–246 (2021)
19. Ehrmann, M., Hamdi, A., Pontes, E.L., Romanello, M., Doucet, A.: Named entity recognition and classification in historical documents: A survey. ACM Computing Surveys **56**(2), 1–47 (2021)
20. Ethayarajh, K.: How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In: Conf. on Empirical Methods in Natural Language Processing, pp. 55–65 (2019)
21. Fornés, A., Romero, V., Baro, A., Toledo, J.I., Sánchez, J., Vidal, E., Lladós, J.: ICDAR2017 competition on information extraction in historical handwritten records. In: Int. Conf. on Document Analysis and Recognition, pp. 1389–1394 (2017)
22. Giotis, A.P., Sfikas, G., Gatos, B., Nikou, C.: A survey of document image word spotting techniques. Pattern Recognition **68**, 310–332 (2017)
23. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. Int. Journal of Computer Vision **129**(6), 1789–1819 (2021)

24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Int. Conf. on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
25. Heinzerling, B., Strube, M.: BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In: Int. Conf. on Language Resources and Evaluation (2018)
26. 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 Conf. on Pattern Recognition, pp. 459–472 (2018)
27. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: OCR-free document understanding transformer. In: European Conf. on Computer Vision, pp. 498–517 (2022)
28. Krishnan, P., Dutta, K., Jawahar, C.V.: HWNet v3: A joint embedding framework for recognition and retrieval of handwritten text. Int. Journal on Document Analysis and Recognition pp. 1–17 (2023)
29. Krishnan, P., Jawahar, C.V.: Bringing semantics in word image retrieval. In: Int. Conf. on Document Analysis and Recognition, pp. 733–737 (2013)
30. Krishnan, P., Jawahar, C.V.: Bringing semantics into word image representation. Pattern Recognition **108**, 107,542 (2020)
31. Landeghem, J.V., Tito, R., Borchmann, L., Pietruszka, M., Jurkiewicz, D., Powalski, R., Józiak, P., Biswas, S., Coustaty, M., Stanislawek, T.: ICDAR 2023 competition on document understanding of everything (DUDE). In: Int. Conf. on Document Analysis and Recognition, pp. 420–434 (2023)
32. Liu, S., Zhang, X., Zhang, S., Wang, H., Zhang, W.: Neural machine reading comprehension: Methods and trends. Applied Sciences **9**(18), 3698 (2019)
33. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019)
34. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press (2008)
35. Marti, U., Bunke, H.: The IAM-database: An English sentence database for offline handwriting recognition. Int. Journal on Document Analysis and Recognition **5**(1), 39–46 (2002)
36. Mathew, M., Gómez, L., Karatzas, D., Jawahar, C.V.: Asking questions on handwritten document collections. Int. Journal on Document Analysis and Recognition **24**, 235–249 (2021)
37. Mathew, M., Karatzas, D., Jawahar, C.V.: DocVQA: A dataset for VQA on document images. In: IEEE Winter Conf. on Applications of Computer Vision, pp. 2199–2208 (2021)
38. Mathew, M., Tito, R., Karatzas, D., Manmatha, R., Jawahar, C.V.: Document visual question answering challenge 2020. CoRR **abs/2008.08899** (2020)
39. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Int. Conf. on Learning Representations (2013)
40. Monroc, C.B., Miret, B., Bonhomme, M., Kermorvant, C.: A comprehensive study of open-source libraries for named entity recognition on handwritten historical documents. In: Int. Workshop on Document Analysis Systems, pp. 429–444 (2022)
41. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Annual Conf. of the North American Chapter of the Association for Computational Linguistics, pp. 2227–2237 (2018)
42. Powalski, R., Borchmann, L., Jurkiewicz, D., Dwojak, T., Pietruszka, M., Palka, G.: Going Full-TILT boogie on document understanding with text-image-layout transformer. In: Int. Conf. on Document Analysis and Recognition, pp. 732–747 (2021)
43. Prasad, A., Déjean, H., Meunier, J., Weidemann, M., Michael, J., Leifert, G.: Bench-marking information extraction in semi-structured historical handwritten records. CoRR **abs/1807.06270** (2018)
44. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100, 000+ questions for machine comprehension of text. In: Conf. on Empirical Methods in Natural Language Processing, pp. 2383–2392 (2016)
45. Rath, T.M., Manmatha, R.: Word spotting for historical documents. Int. Journal on Document Analysis and Recognition **9**(2-4), 139–152 (2007)
46. Rouhou, A.C., Dhiaf, M., Kessentini, Y., Salem, S.B.: Transformer-based approach for joint handwriting and named entity recognition in historical document. Pattern Recognition, Letters **155**, 128–134 (2022)
47. Rowtula, V., Krishnan, P., Jawahar, C.V.: PoS tagging and named entity recognition on handwritten documents. In: Int. Conf. on Natural Language Processing (2018)
48. Rowtula, V., Oota, S.R., Jawahar, C.V.: Towards automated evaluation of handwritten assessments. In: Int. Conf. on Document Analysis and Recognition, pp. 426–433 (2019)
49. Sauer, A., Asaadi, S., Küch, F.: Knowledge distillation meets few-shot learning: An approach for few-shot intent classification within and across domains. In: Workshop on NLP for Conversational AI, pp. 108–119 (2022)
50. Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. In: Int. Conf. on Learning Representations (2017)
51. Sezerer, E., Tekir, S.: A survey on neural word embeddings. CoRR **abs/2110.01804** (2021)
52. Sharma, A., Jayagopi, D.B.: Automated grading of handwritten essays. In: Int. Conf. on Frontiers in Handwriting Recognition, pp. 279–284 (2018)
53. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Int. Conf. on Learning Representations (2015)
54. van Strien., D., Beelen., K., Ardanuy., M.C., Hosseini., K., McGillivray., B., Colavizza., G.: Assessing the impact of OCR quality on downstream NLP tasks. In: Int. Conf. on Agents and Artificial Intelligence, pp. 484–496 (2020)
55. Sudholt, S.: Learning attribute representations with deep convolutional neural networks for word spotting. Ph.D. thesis, TU Dortmund (2018)
56. Sudholt, S., Fink, G.A.: PHOCNet: A deep convolutional neural network for word spotting in handwritten documents. In: Int. Conf. on Frontiers in Handwriting Recognition, pp. 277—282 (2016)
57. Sudholt, S., Fink, G.A.: Evaluating word string embeddings and loss functions for CNN-based word spotting. In: Int. Conf. on Document Analysis and Recognition, pp. 493–498 (2017)
58. Tang, L., Kender, J.R.: Educational video understanding: Mapping handwritten text to textbook chapters. In: Int. Conf. on Document Analysis and Recognition, pp. 919–923 (2005)
59. Tarride, S., Boillet, M., Kermorvant, C.: Key-value information extraction from full handwritten pages. In: Int. Conf. on Document Analysis and Recognition, pp. 185–204 (2023)

60. Tarride, S., Lemaitre, A., Coüasnon, B., Tardivel, S.: A comparative study of information extraction strategies using an attention-based neural network. In: Int. Workshop on Document Analysis Systems, pp. 644–658 (2022)
61. Tito, R., Mathew, M., Jawahar, C.V., Valveny, E., Karatzas, D.: ICDAR 2021 competition on document visual question answering. CoRR **abs/2111.05547** (2021)
62. Toledo, J.I., Carbonell, M., Fornés, A., Lladós, J.: Information extraction from historical handwritten document images with a context-aware neural model. Pattern Recognition **86**, 27–36 (2019)
63. Tüselmann, O., Brandenbusch, K., Chen, M., Fink, G.A.: A weighted combination of semantic and syntatic word image representations. In: Int. Conf. on Frontiers in Handwriting Recognition, pp. 285–299 (2022)
64. Tüselmann, O., Fink, G.A.: Exploring semantic word representations for recognition-free NLP on handwritten document images. In: Int. Conf. on Document Analysis and Recognition, pp. 85–100 (2023)
65. Tüselmann, O., Wolf, F., Fink, G.A.: Identifying and tackling key challenges in semantic word spotting. In: Int. Conf. on Frontiers in Handwriting Recognition, pp. 55–60 (2020)
66. Tüselmann, O., Wolf, F., Fink, G.A.: Are end-to-end systems really necessary for NER on handwritten document images? In: Int. Conf. on Document Analysis and Recognition, pp. 808–822 (2021)
67. Tüselmann, O., Müller, F., Wolf, F., Fink, G.A.: Recognition-free Question Answering on Handwritten Document Collections. In: Int. Conf. on Frontiers in Handwriting Recognition, pp. 259–273 (2022)
68. Villanova-Aparisi, D., Martinez-Hinarejos, C.D., Romero, V., Pastor-Gadea, M.: Evaluation of different tagging schemes for named entity recognition in handwritten documents. In: Int. Conf. on Document Analysis and Recognition, pp. 3–16 (2023)
69. Wang, W., Bi, B., Yan, M., Wu, C., Xia, J., Bao, Z., Peng, L., Si, L.: StructBERT: Incorporating language structures into pre-training for deep language understanding. In: Int. Conf. on Learning Representations (2020)
70. Wilkinson, T., Brun, A.: Semantic and verbatim word spotting using deep neural networks. In: Int. Conf. on Frontiers in Handwriting Recognition, pp. 307–312 (2016)
71. Wolf, F., Fink, G.A.: Self-training of handwritten word recognition for synthetic-to-real adaptation. In: Int. Conf. on Pattern Recognition, pp. 3885–3892 (2022)
72. Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A.R., van den Hengel, A.: Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding **163**, 21–40 (2017)
73. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florêncio, D.A.F., Zhang, C., Che, W., Zhang, M., Zhou, L.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In: Annual Meeting of the Association for Computational Linguistics and Int. Joint Conf. on Natural Language Processing, pp. 2579–2591 (2021)
74. Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models. In: Int. Conf. on Computational Linguistics, pp. 2145–2158 (2018)
75. Yamada, I., Asai, A., Shindo, H., Takeda, H., Matsumoto, Y.: LUKE: Deep contextualized entity representations with entity-aware self-attention. In: Conf. on Empirical Methods in Natural Language Processing, pp. 6442–6454 (2020)
76. Zeng, C., Li, S., Li, Q., Hu, J., Hu, J.: A survey on machine reading comprehension: Tasks, evaluation metrics, and benchmark datasets. Applied Sciences **10**(21), 7640 (2020)
77. Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., Chua, T.: Retrieving and reading: A comprehensive survey on open-domain question answering. CoRR **abs/2101.00774** (2021)