

# Fetch-A-Set: A Large-Scale OCR-Free Benchmark for Historical Document Retrieval

Adrià Molina<sup>1,2\*</sup>[0000-0003-0167-8756], Oriol Ramos Terrades<sup>1,2</sup>[0000-0002-3333-8812], and Josep Lladós<sup>1,2</sup>[0000-0002-4533-4739]

<sup>1</sup> Computer Vision Center

<sup>2</sup> Computer Science Department

Universitat Autònoma de Barcelona, Catalunya

{amolina, oriolrt, josep}@cvc.uab.cat

**Abstract.** This paper introduces Fetch-A-Set (FAS), a comprehensive benchmark tailored for legislative historical document analysis systems, addressing the challenges of large-scale document retrieval in historical contexts. The benchmark comprises a vast repository of documents dating back to the XVII century, serving both as a training resource and an evaluation benchmark for retrieval systems. It fills a critical gap in the literature by focusing on complex extractive tasks within the domain of cultural heritage. The proposed benchmark tackles the multi-faceted problem of historical document analysis, including text-to-image retrieval for queries and image-to-text topic extraction from document fragments, all while accommodating varying levels of document legibility. This benchmark aims to spur advancements in the field by providing baselines and data for the development and evaluation of robust historical document retrieval systems, particularly in scenarios characterized by wide historical spectrum.

**Keywords:** Document Retrieval · Information Extraction · Historical documents · Datasets · Legislative Documents

## 1 Introduction

The automation of document understanding procedures is a growing trend across various industries. Document management systems to extract information, index, summarize or assist in decision making tasks are more and more frequent in fintech, legaltech, insurancetech, among other. A particular case are the systems for smart digitization of historical documents in archives and libraries. With historical data gaining significance in governmental bodies, heritage management is not exempt from this shift towards automation.

---

\* Main Corresponding Author

With the growing of Document Intelligence, large scale digitization processes of historical documents for digital preservation purposes have scaled up to systems that are able to understand the contents providing innovative services to different communities of users. In this paper we consider two challenges faced by heritage institutions in handling vast documental sources. Firstly, from the user perspective, we tackle the need for continuous indexing of databases to enable natural language queries as a “text-to-image” task. In other words, we aim to fetch relevant documents (image) based on human-written queries (text). We refer to this semantic-based task as *topic spotting* (Figure 1, left), differentiating it from the traditional word spotting that looks for exact matchings between the query words and the document content. Secondly, on the institutional front, managing large databases with millions of historical records becomes impractical for humans, causing delays in public access. Recognizing the importance of categorizing historical data in archival procedures, our paper also explores the “image-to-text” task. This task, namely *information extraction* (Figure 1, right), seeks to provide a feasible set of texts from an image, aiding users in establishing prior knowledge automatically for incoming data. Therefore, incorporating complex understanding tasks in the realm of historical document analysis systems will result in novel services to understand the history.

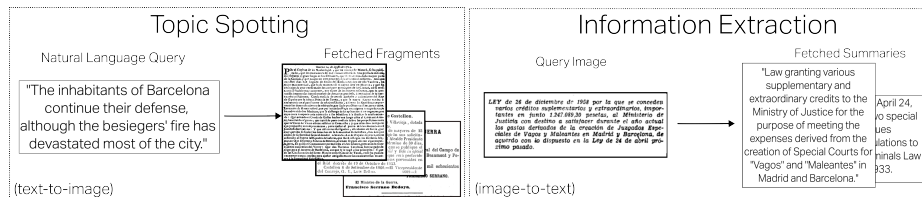


Fig. 1: Illustration depicting the primary objectives pursued by FAS in evaluating and training information extraction systems. Specifically, retrieving document fragments based on a given topic in natural language (right) or generating plausible descriptions from a given fragment (left).

The inherent nature of historical documentation renders the reliance solely on text-based approaches for information retrieval unfeasible [15]. The damaged condition of the documents, the lack of accurate ground truth, and the complexities of dealing with multiple languages severely impede the effectiveness of text-based methods. Consequently, it becomes crucial to incorporate visual insights to overcome these limitations and enable topic-sensitive retrieval in historical document analysis. Note that the existing historical document benchmarks [25] mainly concentrate on tasks such as classification [17], date and writer retrieval [5], or word spotting [20], but they do not specifically address the challenges associated with topic-aware document retrieval. This scarcity of dedicated datasets underscores the need for research that explores document understanding in historical contexts.

This work proposes Fetch-A-Set (**FAS**) as a benchmark to evaluate the effectiveness of document understanding systems in fetching relevant information directly from natural text, eliminating the need for expensive OCR solutions, especially for large historical collections with significant temporal variance.

To demonstrate the relevance of this evaluation to the community, we summarize the contributions of this work as follows:

- We introduce FAS as a benchmark for evaluating content-based extractive systems on historical records. FAS significantly enhances data availability, encompassing 400K samples sourced from three centuries of Spanish legislative documents (Section 3).
- To demonstrate FAS’s significance and its potential to inspire further research, we present a classical OCR-based baseline alongside a vision-based approach. We illustrate the contexts in which each method excels, providing guidance to authors for conducting more effective comparisons in their methodologies (Sections 4, 5).
- Finally, in Section 6, we explore how historical bias may contribute to improve performance in low-legibility scenarios.

## 2 Related Work

As introduced in Section 1, the availability of OCR annotations in historical data is usually scarce. Moreover, recent analysis conducted by Hamdi *et al.* (2023) [15] reveals that document extractive tasks achieve an accuracy of 80-90% when applied directly to textual data benchmarks. However, in the same study, it is noted that integrating a visual perspective requires character recognition, which tends to drop the performance as the recognition systems are not adapted for such data. In historical document analysis, noise in OCR predictions poses a common challenge. In this context, retrieval tasks aim to directly index content from the visual domain. However, this section will explore how the historical document analysis community has predominantly focused on word spotting, driven by the necessity upon which this work is based: indexing text in large databases.

In fact, Nikolaidiou et al. (2022) [25] conducted an extensive survey on document analysis in historical contexts. Out of the 65 datasets they considered, 14 were identified as holding retrieval tasks. These tasks, detailed in Table 1, primarily revolve around word spotting. Notably, 10 to 13 datasets focus on this task, with “IAM-HistDB” often treated as a single entity consisting on [12] and [13]. However, other historical document-related tasks emphasize visual properties over textual patterns. For instance, some tasks involve estimating dates from scanned photographs [22,24,36] or retrieving iconography/ornaments [9] from medieval documents. Additionally, certain tasks, like writer identification

Dataset (HDR)	Retrieval	#Samples ( $\sim$ )	Time Span	Type	Script
GERMANA <sub>2009</sub> [28]	WS	200k	1891	HW	Latin
RODRIGO <sub>2010</sub> [33]	WS	200K	1545	HW	Latin
Saint-Gall <sub>2011</sub> [12]	WS	5K	IX c.	HW	Latin
Parzival <sub>2012</sub> [13]	WS	20K	XIII c.	HW	Gothic
Washington <sub>2012</sub> [13]	WS	5K	XVIII c.	HW	Latin
Esposalles <sub>2013</sub> [31]	WS	2K	1451 - 1495	HW	Latin
BH2M <sub>2014</sub> [10]	WS	50K	1617 - 1619	HW	Latin
HADARA80P <sub>2014</sub> [27]	WS	15K	1430	HW	Arabic
ENP <sub>2015</sub> [7]	WS	-	XVIII - XX c.	Printed	Latin
GRPPLY-DB <sub>2015</sub> [14]	WS	100k	1838 - 1912	Mixed	Greek
AMADI <sub>2016</sub> [19]	WS	10K	XV c.	Lontar	Hanacaraka
VML-HD <sub>2017</sub> [18]	WS	200K	1088-1451	HW	Arabic
CFRAMUZ <sub>2017</sub> [2]	WS	20K	1910-1946	HW	Latin
DocExplore <sub>2016</sub> [9]	Ornament	2K	X - XVI c.	Image	-
ICDAR17 H-WI <sub>2017</sub> [11]	Writer	5K	XIII - XX c.	HW	Latin
ICDAR19-HDRC-IR <sub>2019</sub> [5]	Writer	20K	IX - XVII c.	HW	Latin
Papy-Row <sub>2021</sub> [6]	Writer	6K	VI c.	Papyrus	Greek
HistDIA <sub>2021</sub> [34]	Date, Font, Loc	10/30/5 K	IX-XVIII c.	HW	Latin
DEW <sub>2017</sub> [22]	Date	1M	1930 - 1999	Image	-
IMAGO <sub>2020</sub> [36]	Date	80K	1845 - 2009	Image	-
DEW-B <sub>2024</sub> [24]	Date	1.5M	1930 - 1999	Image	-

Table 1: Table for most popular historical document retrieval datasets, "retrieval" stands for the retrieved content in the task: Word Spotting (**WS**), **Ornament**, **Writer** and **Font** Identification and **Date** and **Location** estimation.

[5,6,11], tackle subtle intra-class differences (visual attributes) alongside exact pattern processing (recognition).

Historical document retrieval literature commonly emphasizes textual content characterization (Table 1), particularly through word spotting and writer identification tasks. This preference indicates a significant interest in deciphering the written content of documents.

Consequently, the upcoming sections will introduce the **FAS** benchmark. This dataset serves as an endeavor to assess systems designed to extract meaning from historical documents, moving beyond mere word spotting and delving into the comprehension of textual content from the vision perspective.



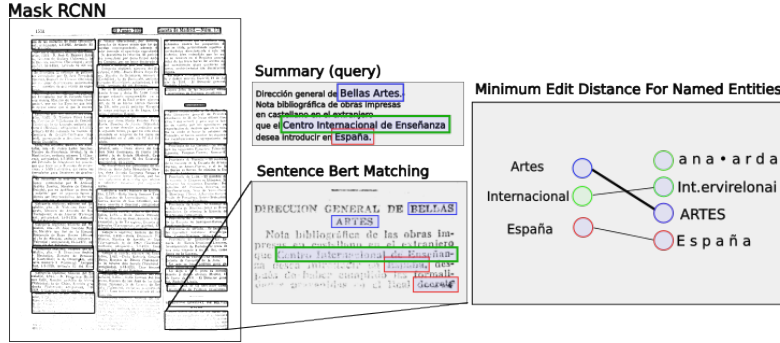


Fig. 2: Query-Region Matching system for creating the Ground Truth (GT) for FAS dataset. Since the presence of the fragment  $F$  in the document  $d$  is human-annotated, and the number of invalid fragments within a page is typically low, the risk of adding noise is significantly reduced.

### 3 The Fetch-A-Set Dataset

In this section, the construction and basic analytics of the **FAS** dataset are presented. This benchmark comprises a set of full-page documents with an associated natural language query. We design a solid heuristic for assigning a one-to-one relationship between fragments of the document and queries. In this way, we expect retrieval systems to be able to replicate this correspondences out of a randomly selected distractor set.

#### 3.1 Building the FAS dataset

Our research extensively utilizes the publicly accessible repository of the “Boletín Oficial del Estado” (BOE), Spain’s Official State Gazette, a pivotal source for disseminating government-approved laws, decrees, provisions, resolutions, and regulations<sup>3</sup>. This repository, accessible online, encompasses historical documents dating back to the XVII century, containing approximately one million pairs of human annotated **summaries** ( $q$ ) and corresponding **documents** ( $d$ ), along with OCR text extracted with a commercial recognition system<sup>4</sup>. As expected and discussed in Section 1, the transcriptions are imperfect. The sentence-bert distance [30] between the **OCR transcription** and the queries ( $q$ ) is considered as measurement for **legibility**; as noisy transcriptions tend gradually lose meaningfulness with respect the original text as the transcription degrades.

<sup>3</sup> <https://www.boe.es/buscar/gazeta.php>

<sup>4</sup> Abbyy Recognition Server v4

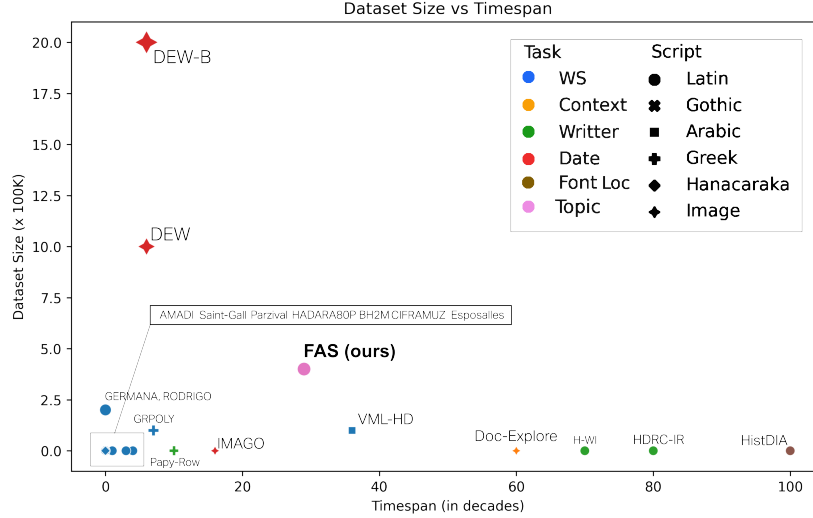


Fig. 3: A scatterplot depicting FAS as the preminent historical document-based retrieval benchmark among those previously examined, while also maintaining a substantial temporal scope. This characteristic enhances the robustness of historical document analysis systems by encompassing wider temporal breadths.

Given the multifaceted nature of newspapers, meaning that (specially in most modern samples) a document contains some irrelevant fragments on the page for a few relevant ones, one of the central challenges in curating this dataset is associating queries  $q$  with specific fragments  $F$  within documents. To address this, we implement a two-step selection process: first, a Mask-RCNN [16] model trained on the Prima Layout Analysis Dataset [1] is used to identify relevant document regions, followed by matching queries with OCR text using sentencebert encoding [30]. Notably, the OCR transcriptions may be unreliable, and to enhance the accuracy of query-document associations, a filtering step employing the Hungarian algorithm [21] with edit distance measure is used to assess the similarity of named entities<sup>5</sup> as illustrated in Figure 2.

In adopting this approach, we have established a robust foundation for ensuring the accuracy of information extraction (**image-to-text**) and topic spotting (**text-to-image**). To manage practical and computational constraints, we initially focused on a well-curated subset of 400K **fragments of documents**<sup>6</sup>. As seen in 2, we randomly divide the given set of documents in a **train** and **test** split, with an additional set of 1024 **distractor** documents that will serve as an

<sup>5</sup> Extracted with SpaCy <https://spacy.io/models/es>.

<sup>6</sup> Train / Test / Distractor splits with queries, hyper-references to the documents (whole page), regions of interest and OCR transcriptions can be found at <http://datasets.cvc.uab.es/B0Ev2/B0Ev2.zip>.

evaluation anchor for retrieval in order to avoid expensive computations and the usage of the whole training set as a distractor, which, due to the challenges of the task, would drop the performance to non-significant comparisons.

#Train	#Test	#Distractors
384567	42997	1024

Table 2: Table showing train, test and distractor number of samples in FAS.

### 3.2 Dataset Analytics

As illustrated in Figure 3 and detailed in Table 1, the majority of retrieval datasets primarily emphasize word spotting tasks. However, it is worth noting a significant bias towards datasets with narrower time spans, with the exception of VML-HD [18]. Despite this bias, certain benchmarks focusing on textual analysis, such as [5,11,34], extend the temporal range up to 1000 years.

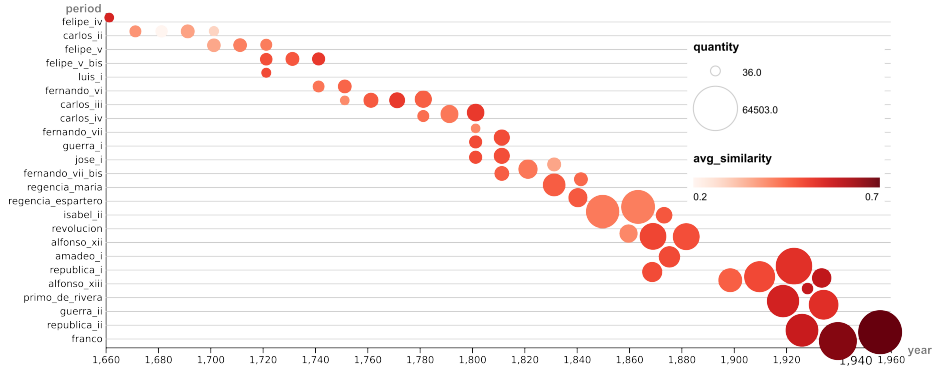


Fig. 4: Beeswarm plot showing the time arrow of the dataset with the years (X-Axis), historical period (Y-Axis) the quantity of documents (size) and the average legibility score (hue).

While the pursuit of temporal diversity may seem like a matter of mere curiosity to the general audience, in certain contexts, it holds significant relevance. Specifically, the consideration of systems that continuously ingest historical documents from various sources can offer valuable insights. In the realm of historical document analysis, methods often prioritize robustness by focusing narrowly on specific applications (see Figure 3), neglecting the potential benefits of incorporating temporal diversity. In this regard, the FAS framework emerges as a

pivotal convergence point between variance and scale. It addresses the need for extractive textual tasks in retrieval systems, positioned at the top of scale in text analysis while encompassing a significant time span.

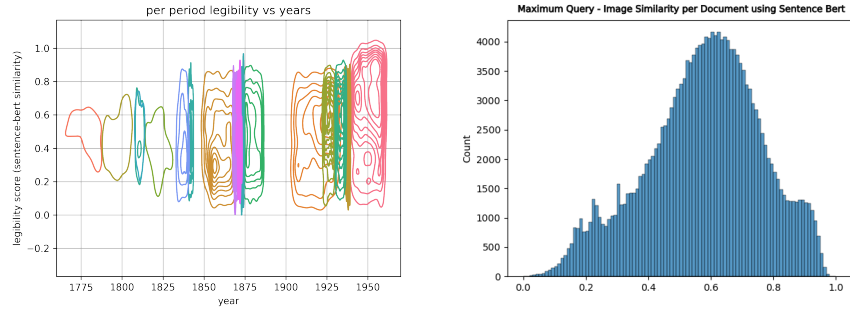


Fig. 5: Left: Distribution of legibility (Y-Axis) through the years (X-Axis) per historical period (hue). Right: Global distribution of legibility score.

As shown in Figure 4, around 100k samples from the FAS dataset belong to the period of time of Franco’s fascist dictatorship over Spain (1939-1975, limited to 1959 by BOE’s organization); followed (80K samples) by Alfonso XIII’s reign (1902-1923 / 1931). This unbalance, as a consequence of both time variance of periods and increasing contemporaneity of the events, is furtherly explored in Figure 5; where we observe a clearer correlation between legibility and years.

As previously noted in Figure 4, it seems that average legibility is expressed through the years. While it may become apparent that legibility increases with time, it is an effect of averaging scores. In Figure 5 we conclude that high legibility has a bias towards modernity, but not the other way around, as many samples in most modern documents also contain poor legibility scores. As variance in deployed document retrieval systems is a mandatory property when pursuing robustness, this captured variance in most modern documents shall promote dataset’s strength in contributing to the historical document analysis community.

As depicted in Figure 6, the extensive temporal breadth results in considerable variability in layout. This variability holds promise for the development of systems intended to accurately capture such contextual variance in their application domain.

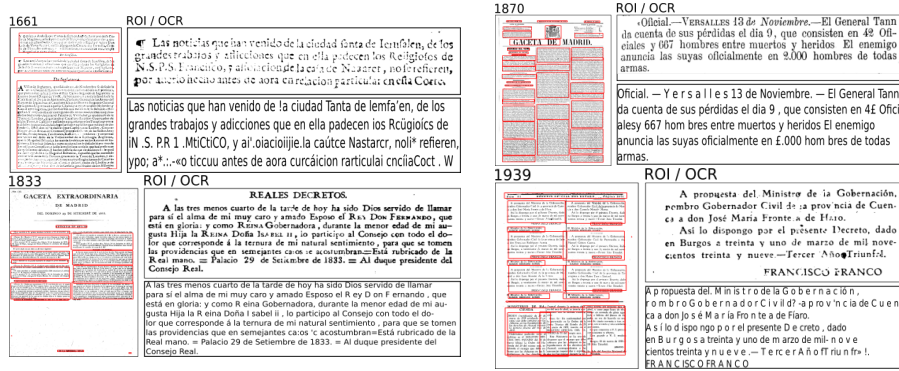


Fig. 6: Examples of segmented documents (1661, 1833, 1870 and 1939) among its corresponding ROI and extracted OCR.

## 4 Baselines

As previously introduced, the primary objective of this study is to tackle the challenge of evaluating information extraction systems within a retrieval framework. We formulate a multi-modal retrieval problem that can be perceived as both information extraction (image-to-text) or topic spotting (text-to-image). In this section, we propose two different baselines serving as pivot to measure the performance of further developed retrieval systems. To achieve this, we leverage two distinct views for each document  $d$ : firstly, the query  $q$ , which is encoded using a text encoder  $\tau = \phi_t(q)$ , and secondly, the image representing the corresponding fragment  $F$ , encoded through a visual embedding  $\nu = \phi_v(F)$ . These two distinct representations—textual and visual—allow us to explore information retrieval by bridging the gap between images and natural language queries. Since newspapers usually contain a wide variety of topics for each page, the fragment  $F$  is the region of the document which corresponds to the natural language description  $q$ .

In order to show the relevance of this benchmark to authors and practitioners, two baselines are defined below to test the evaluation capabilities of FAS. As shown in Figure 7, both options are retrieval-based. In the case of the vision-based approach, a fragment is encoded at pixel-level to obtain a close representation to the query. In the case of the OCR-Based approach, we take advantage of the available OCR in the dataset to get an encoded representation of both query and OCR texts.

**Vision-Based Approach** In order to obtain a robust representation of the content at pixel level, we propose the usage of a Vision Transformer [8] (ViT-B/32) with 32x32 patch size pretrained on CLIP [29]. This ViT embeddings are linearly projected and fine-tuned into the *topic space*, which is shared with the

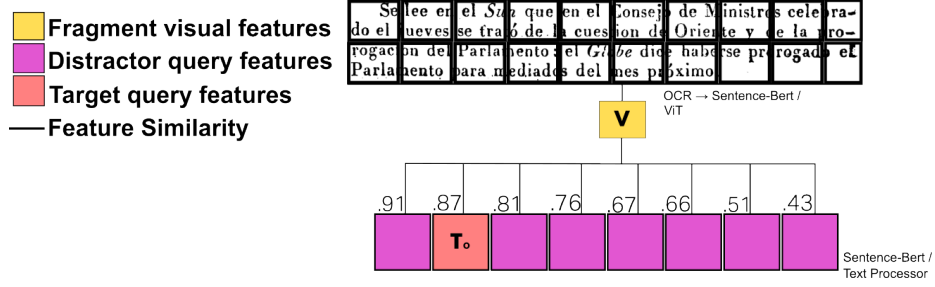


Fig. 7: Basic scheme of the proposed methods where a set of features  $\nu$  are extracted from a fragment. Then, the corresponding query feature  $\tau_0$  is inserted among the set of distractor queries. When comparing distances, the tuple  $(\nu, \tau_0)$  is expected to show maximum similarity. For text-to-image task, the objective is analogous.

query encoder. In doing so, we make use of the standard transformer architecture [35,37,38,39] with 2 layers and 4 heads. We take the decision of training the query embeddings from scratch since the complexity of the textual information present in the dataset does not appear to hold the necessity of bigger architectures. We simultaneously fine-tune the visual encoder and train the query space to be projected at close points in the *topic space*. For doing so, we leverage triplets for using metric learning approaches such as Triplet Margin [3], or contrastive losses such as SimCLR [4,26] and CLIP [29]. In this scenario, the fragment representation  $v$  is encoded through a fine-tuned ViT, while the query  $q$  is processed through a text encoder to obtain a representation  $\tau$ .

**OCR-Based Approach** In order to conduct a comparison with a state-of-the-art text-based approach, we take advantage of the OCR transcriptions acquired with the commercial OCR present in the dataset. We consider the document features as the recognized text processed through the same *sentence-bert*[30] encoder than the queries. We then explore the extend to which a given document matches its corresponding query whenever its distance is minimum among the distractor set.

## 5 Experiments

### 5.1 Objective

The evaluation proposed in the **FAS** benchmark, aims to minimize the distance between the encoded representations of the target document or query  $(\tau, \nu)$  and all elements within the distractor set  $(\mathfrak{Q})$ . Mathematically, the objective (image-to-text) is expressed as  $\min \|\phi_t(q_o) - \phi_v(F_o)\|$  and,  $\forall q_i \in \mathfrak{Q}, \max \|\phi_t(q_i) - \phi_v(F_o)\|$

Approach		Topic Spotting (text-to-image) Metrics					
Architecture	Loss	Acc@1	Acc@5	Acc@10	AvgRank ↓	Acc@1 <sub>low</sub>	Acc@5 <sub>low</sub>
ViT-B/32 [29,8]	Triplet [3]	0.473	0.639	0.695	52.74	0.289	0.419
	SimCLR[4]	0.509	0.672	0.723	44.12	<b>0.296</b>	<b>0.443</b>
	CLIP[29]	0.460	0.657	0.719	32.33	0.264	<b>0.447</b>
OCR + sBert[30]	-	<b>0.657</b>	<b>0.785</b>	<b>0.837</b>	<b>16.46</b>	0.124	0.304

Approach		I. Extraction (image-to-text) Metrics					
Architecture	Loss	Acc@1	Acc@5	Acc@10	AvgRank ↓	Acc@1 <sub>low</sub>	Acc@5 <sub>low</sub>
ViT-B/32 [29,8]	Triplet[3]	0.523	0.640	0.692	64.00	0.326	0.421
	SimCLR[4]	0.522	0.675	0.726	43.05	<b>0.341</b>	0.458
	CLIP[29]	0.482	0.662	0.725	<b>32.99</b>	0.326	<b>0.464</b>
OCR + sBert[30]	-	<b>0.623</b>	<b>0.723</b>	<b>0.760</b>	44.19	0.073	0.149

Table 3: Text-based and Vision-Based baselines for topic spotting (top) and information extraction (bottom). In further analysis we show how visual baseline can outperform OCR+Bert in some challenging scenarios for recognition. The subindex *low* stands for the percentile 0-25% of worst legible documents.

with  $\{q_o\} \cap \mathfrak{Q} = \emptyset$ . This objective aims to identify the query ( $q$ ) within the given distractor set ( $\mathfrak{Q}$ ) that exhibits the minimum distance to the encoded representation of the target fragment ( $F_o$ ), being  $q_o$  the correct correspondence. The encoded representations, derived from the text ( $\phi_t$ ) and visual ( $\phi_v$ ) embeddings, enable the retrieval system to effectively match queries and documents within a multi-modal retrieval framework.

## 5.2 Evaluation Metrics

The previously defined objective, shall be evaluated using top-K Accuracy **Acc@{1, 5, 10}** and, in some experiments, average ranking (**AR**) indicating the average position ( $0, |\mathfrak{Q}|$ ) where the correct matching ( $\tau$  or  $\nu$ ) has been placed with respect the set of distractor data  $\mathfrak{Q}$ .

## 5.3 Results

In Table 3, the quantitative results of both vision and OCR-based approaches are presented. As reported in [23], it is expected to all metric learning approaches

to perform competitively when compared with the other ones, which is the case for the evaluated visual baseline system.

Upon observation, we note that vision approaches prove quantitatively more advantageous than the OCR-based baseline in the context of tasks containing poorly legible documents. On the other hand, using text features becomes significantly more convenient in situations where the recognition can be performed correctly.

The key takeaway from both textual and vision-based baselines is that text features perform exceptionally well when the text is clear and legible. However, there’s potential for vision-based systems in scenarios where visual insights offer an advantage. As discussed in Section 6, the visual benchmark incorporates both keyword spotting and layout-driven date estimation. This implies that hybrid systems capable of leveraging visual features when text is unavailable, and text features otherwise, could enhance the robustness of systems evaluated with this dataset.

## 6 Discussion

From a qualitative standpoint, our primary emphasis is on evaluating both retrieval performance and the comprehensibility of the results. Given the consideration that, as shown in Section 5 Table 3, ViT performance seems to hold even in situations where legibility is low. In this section, we aim to explore which features are being utilized in terms of encoding topic information.

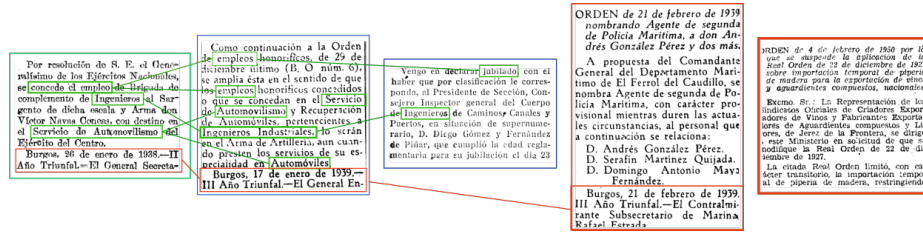


Fig. 8: Success case of the retrieval of documents related to the query “Promotion to the rank of Sergeant in the Complementary Engineer Brigade is hereby granted to Mr. Víctor Navas Concas” (translated). Some relevant semantic features (green), some signs of bias (red) and the relevant document ranked at the top (left, first).



In Figure 8, we observe a document that has been accurately ranked as the top result by our 1-Nearest Neighbor approach. Interestingly, the top-ranked documents contain significant content related to engineering job status. The presence of common keywords serves as a necessary, yet not sufficient, condition for the encoder to perform implicit recognition. However, there are also indications of potential bias in our results. All documents belong to the same period of time, some documents exhibit a shared historical fragment at the bottom (“II Victorious Year”, “III Victorious Year”), reminiscent of the early years of the fascist military uprising in Spain (1936-1939), therefore, it is important to exercise caution when interpreting the encoder’s ability to recognize important words in the text.

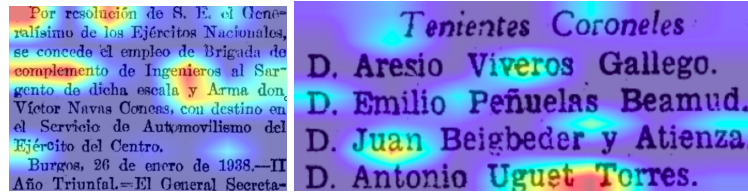


Fig. 9: GradCam [32] activations on fragments  $F$  at the visual tokenizer of the ViT baseline when optimized towards  $q$  alignment.

To investigate whether the visual extractor relies on the text content or if other biases contribute to its performance, we conduct an inspection of the activations ([32]) of the visual tokenizer given its ground truth topic (query,  $q$ ) Figure 9. As it can be observed, there is no sign of bias towards the usage of the bottom fragment of the text, moreover, in Figure 10 it is shown how the visual representation of the fragment  $F$  efficiently tokenizer words to its stems; which should be a sign of robustness to noise and degradation.

Visual Tokenizer Activation	Word	Word Token
	Infanteria	<infanter>
	Cuerpo	<cuerp>
	Administración	<administ>

Fig. 10: Examples of visual words (left) where the visual tokenizer (activations) prioritizes only a patch of the image rather than combining several patches.

However, it is apparent (Figure 8) than the ViT is incorporating some (beneficial) temporal bias; in Figure 11 we ask ourself whether this is happening at text

Layer	Augm <sup>a</sup>					
	None	2x2	4x4	8x8	16x16	Flip <sup>b</sup>
Tok	71.8	79.8	85.4	88.5	118.2	70.3
LN1	63.2	70.2	121.5	79.9	94.3	64.2
Tr <sub>1</sub>	36.8	37.4	36.4	32.9	35.6	37.4
Tr <sub>6</sub>	17.2	18.7	21.7	22.1	24.6	21.5
Tr <sub>12</sub>	14.4	16.1	19.8	22.1	26.1	23.2
LN2	11.6	12.8	15.6	17.9	21.6	18.2
Out	17.2	16.0	19.2	20.1	24.2	20.9

**Tema 86:** De vista de la actividad de la Guardia Municipal de Aduana, sobre Registro de la propiedad de Cuenca por Real Cédula de 16 de Abril último, a fin de que se le expusiera al Regidor de Cuenca, que se le dio el 7 de Mayo de 1682 hasta 9 de Junio de 1683, en que así por tal vez se le pudiese presentar a la Comisión de la misma.

**Tema 87:** De vista de la actividad de la Guardia Municipal de Aduana, sobre Registro de la propiedad de Cuenca por Real Cédula de 16 de Abril último, a fin de que se le expusiera al Regidor de Cuenca, que se le dio el 7 de Mayo de 1682 hasta 9 de Junio de 1683, en que así por tal vez se le pudiese presentar a la Comisión de la misma.

**Tema 86:** De vista de la actividad de la Guardia Municipal de Aduana, sobre Registro de la propiedad de Cuenca por Real Cédula de 16 de Abril último, a fin de que se le expusiera al Regidor de Cuenca, que se le dio el 7 de Mayo de 1682 hasta 9 de Junio de 1683, en que así por tal vez se le pudiese presentar a la Comisión de la misma.

**Tema 87:** De vista de la actividad de la Guardia Municipal de Aduana, sobre Registro de la propiedad de Cuenca por Real Cédula de 16 de Abril último, a fin de que se le expusiera al Regidor de Cuenca, que se le dio el 7 de Mayo de 1682 hasta 9 de Junio de 1683, en que así por tal vez se le pudiese presentar a la Comisión de la misma.

<sup>a</sup> NxN: Average blurring  $K_{size} = N$ .  
<sup>b</sup> Flip: Horizontal and vertical flipping

Fig. 11: Mean Absolute Error on date estimation given the internal ViT representation of documents (tokenizer, normalization, transformer and output) given different sizes of blurring and examples on incrementally blurred documents.

level (which would contradict Figure 9) or ViT is focusing on some other features to incorporate temporal scale in the process of assigning a topic representation to a given image.

As illustrated in Figure 11, it is evident that the deeper the visual representation utilized, the more accurate the date estimation by a linear regression. This effect is substantial, with average errors ranging from 11 to 15 years in some cases. Moreover, this trend persists even with increasingly heavier degradations of the fragment, indicating the sustained performance of the model.

From these findings, we can infer that there is indeed temporal information embedded within the tokens that define the topic space (as depicted in Figure 8). However, while there is also sensitivity to text (as shown in Figures 9 and 10), the date estimation primarily occurs at the layout level rather than by associating specific words with the temporal scale (Figure 11).

## 7 Conclusions

This paper has presented a novel dataset and baselines for evaluating historical document retrieval systems, termed Fetch-A-Set (FAS), which proves effective in assessing the performance of both text and vision-based systems. We underscore the necessity of transitioning from individual words to natural text to enhance historical document management in terms of indexing and comprehension, and we provide authors with a comprehensive dataset for training and evaluation purposes.

Furthermore, the paper stresses the importance of employing large-scale datasets for training and evaluation, as they encompass a broader temporal variance relevant to the application of systems developed by the community.

Lastly, we evaluate both baselines and investigate the impact of vision on the results. Our findings suggest that incorporating temporal bias into retrieval systems is advantageous, even without relying on text features. This exploration suggests that both vision and text are niche solutions for solving FAS, which should help to evaluate text, vision-based systems and hybridisation systems that aim to provide robustness by incorporating both modalities into the system.

## Acknowledgment

This work has been partially supported by the Spanish project PID2021-126808OB-I00, Ministerio de Ciencia e Innovación, the Departament de Cultura of the Generalitat de Catalunya, and the CERCA Program / Generalitat de Catalunya. Adrià Molina is funded with the PRE2022-101575 grant provided by MCIN / AEI / 10.13039 / 501100011033 and by the European Social Fund (FSE+).

## References

1. Antonacopoulos, A., Bridson, D., Papadopoulos, C., Pletschacher, S.: A realistic dataset for performance evaluation of document layout analysis. In: 2009 10th International Conference on Document Analysis and Recognition. IEEE
2. Arvanitopoulos, N., Chevassus, G., Maggetti, D., et al.: A handwritten french dataset for word spotting: Cframuz. In: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing (2017)
3. Balntas, V., Riba, E., Ponsa, D., Mikolajczyk, K.: Learning local feature descriptors with triplets and shallow convolutional neural networks. In: Bmvc
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning
5. Christlein, V., Nicolaou, A., Seuret, M., Stutzmann, D., Maier, A.: Icdar 2019 competition on image retrieval for historical handwritten documents
6. Cilia, N.D., De Stefano, C., Fontanella, F., Marthot-Santaniello, I., Scotto di Freca, A.: Papyrow: a dataset of row images from ancient greek papyri for writers identification. In: International Conference on Pattern Recognition. pp. 223–234. Springer (2021)
7. Clausner, C., Papadopoulos, C., Pletschacher, S., Antonacopoulos, A.: The enp image and ground truth dataset of historical newspapers. pp. 931–935 (08 2015). <https://doi.org/10.1109/ICDAR.2015.7333898>
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (2020)

9. En, S., Nicolas, S., Petitjean, C., Jurie, F., Heutte, L.: New public dataset for spotting patterns in medieval document images. *Journal of Electronic Imaging* **26**(1), 011010 (2016). <https://doi.org/10.1117/1.JEI.26.1.011010>, <https://doi.org/10.1117/1.JEI.26.1.011010>
10. Fernández-Mota, D., Almazán, J., Cirera, N., Fornés, A., Lladós, J.: Bh2m: The barcelona historical, handwritten marriages database. In: 2014 22nd International Conference on Pattern Recognition. pp. 256–261. IEEE (2014)
11. Fiel, S., Kleber, F., Diem, M., et al.: Icdar2017 competition on historical document writer identification (historical-wi). In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). pp. 1377–1382 (2017)
12. Fischer, A., Frinken, V., Fornés, A., Bunke, H.: Transcription alignment of latin manuscripts using hidden markov models. In: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. pp. 29–36 (2011)
13. Fischer, A., Keller, A., Frinken, V., Bunke, H.: Lexicon-free handwritten word spotting using character hmms. *Pattern recognition letters* **33**(7), 934–942 (2012)
14. Gatos, B., Stamatopoulos, N., Louloudis, G., Sfikas, G., Retsinas, G., Papavassiliou, V., Sunistira, F., Katsouros, V.: Grpoly-db: An old greek polytonic document image database. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 646–650 (2015). <https://doi.org/10.1109/ICDAR.2015.7333841>
15. Hamdi, A., Pontes, E.L., Sidere, N., Coustaty, M., Doucet, A.: In-depth analysis of the impact of ocr errors on named entity recognition and linking. *Natural Language Engineering* **29**(2), 425–448 (2023)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision
17. Jaume, G., Ekenel, H.K., Thiran, J.P.: Funsd: A dataset for form understanding in noisy scanned documents. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). IEEE
18. Kassiss, M., Abdalhaleem, A., Droby, A., et al.: Vml-hd: The historical arabic documents dataset for recognition systems. In: 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR). pp. 11–14 (2017)
19. Kesiman, M., Burie, J., Wibawantara, G., et al.: Amadi\_lontarset: The first handwritten balinese palm leaf manuscripts dataset. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 168–173 (2016)
20. Krishnan, P., Jawahar, C.: Hwnet v2: an efficient word image representation for handwritten documents. *International Journal on Document Analysis and Recognition (IJDAR)*
21. Kuhn, H.W.: The hungarian method for the assignment problem. *Naval research logistics quarterly* **2**(1-2), 83–97 (1955)
22. Müller, E., Springstein, M., Ewerth, R.: “when was this picture taken?”—image date estimation in the wild. In: Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings 39. pp. 619–625. Springer (2017)
23. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. In: Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16
24. Net, F., Hernández, N., Molina, A., Gómez, L.: A transformer-based object-centric approach for date estimation of historical photographs. In: European Conference on Information Retrieval. pp. 137–150. Springer (2024)
25. Nikolaidou, K., Seuret, M., Mokayed, H., Liwicki, M.: A survey of historical document image datasets (2022)

26. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2018)
27. Pantke, W., Dennhardt, M., Fecker, D., Märgner, V., Fingscheidt, T.: An historical handwritten arabic dataset for segmentation-free word spotting - hadara80p. In: 2014 14th International Conference on Frontiers in Handwriting Recognition. pp. 15–20 (2014). <https://doi.org/10.1109/ICFHR.2014.11>
28. Pérez, D., Tarazón, L., Serrano, N., et al.: The germana database. In: 2009 10th International Conference on Document Analysis and Recognition. pp. 301–305 (2009)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR
30. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks
31. Romero, V., Fornés, A., Serrano, N., Sánchez, J.A., Toselli, A.H., Frinken, V., Vidal, E., Lladós, J.: The esposalles database: An ancient marriage license corpus for off-line handwriting recognition. *Pattern Recognition* **46**(6), 1658–1669 (2013)
32. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
33. Serrano, N., Castro, F., Juan, A.: The rodrigo database. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10). European Language Resources Association (ELRA), Valletta, Malta (2010), [URL](#)
34. Seuret, M., Nicolaou, A., Rodríguez-Salas, D., Weichselbaumer, N., Stutzmann, D., Mayr, M., Maier, A., Christlein, V.: Icdar 2021 competition on historical document classification. In: International Conference on Document Analysis and Recognition. pp. 618–634. Springer (2021)
35. Shazeer, N.: Glue variants improve transformer. arXiv preprint [arXiv:2002.05202](#) (2020)
36. Stacchio, L., Angeli, A., Lisanti, G., Calanca, D., Marfia, G.: Imago: A family photo album dataset for a socio-historical analysis of the twentieth century. arXiv preprint [arXiv:2012.01955](#) (2020)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems*
38. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: International Conference on Machine Learning. pp. 10524–10533. PMLR (2020)
39. Zhang, B., Sennrich, R.: Root mean square layer normalization. *Advances in Neural Information Processing Systems* **32** (2019)