HistoSketch: A Semi-Automatic Annotation Tool for Archival Documents

Joan Mas, José A. Rodríguez, Dimosthenis Karatzas, Gemma Sánchez, Josep Lladós
Computer Vision Center - Computer Science Department
Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain
{jmas,jrodriguez,dimos,gemma,josep}@cvc.uab.cat

Abstract

This article describes a sketch-based framework for semi-automatic annotation of historical document collections. It is motivated by the fact that fully automatic methods, while helpful for extracting metadata from large collections, have two main drawbacks in a real-world application: (i) they are error-prone and manual intervention is always required, and (ii) they only capture a subset of all the knowledge in the document base. Therefore, we have developed a practical framework for allowing experts to extract knowledge from document collections in a sketch-based scenario. The main possibilities of the proposed framework are: (a) browsing the collection efficiently, (b) providing gestures for metadata input, (c) supporting handwritten notes and (d) providing gestures for launching automatic extraction processes such as OCR or word spotting.

1 Introduction

In recent years, we have witnessed an increasing activity in the field of digital libraries [1, 3, 17]. Among the document image analysis (DIA) community, a special interest has been put on historical documents. The digitalization of such kind of documents not only allows the preservation of the originals, but also makes them publicly available through the internet.

Especially when the amount of data increases, it is necessary to provide metadata associated to the documents for indexing and retrieval purposes, for allowing easy browsing, queries, efficient visualization, etc. To avoid manual intervention, some classical techniques such as OCR [4] or word spotting [13] can be applied. However, these techniques might not be complete in the sense that (i) they only provide a part of the interesting metadata, and (ii) as other pattern recognition methods, they are error-prone. This means that an expert is always required at a final step, either to correct the output of the automatic algorithms or to introduce new knowledge not captured by it, for instance lexical relationships (not considered by an OCR).

Therefore, we believe that it is important to complement the automatic tools for metadata extraction with tools that provide an efficient human-computer interface to experts. Such tools could allow to manually introduce new metadata or annotations, start automated processes, correct the output of those, etc.

Although many researchers have exploited the possibilities that sketch-based interfaces offer in different application frameworks, such as user interface design [9] and diagram understanding [7], very few works facing the problem of annotation of historical documents may be found.

In this work, we present a sketch-based interface that allows experts to manually generate metadata in different ways and that complements other automatic metadata extraction techniques. More precisely, the sketch-based interface provides the means for browsing the collection, marking important elements, establishing some relationships between the elements and introducing handwritten annotations. The archive of border records from the Civil Government of Girona, Spain is used here as a test-bed to evaluate the effectiveness of this approach.

The rest of the article is structured as follows. Section 2 provides a description of the specific collection and the metadata of interest. Section 3 explains the automatic DIA information extraction techniques that are already considered. Section 4 describes in detail the main focus of this article: the sketch-based interface for knowledge capture and extraction. Section 5 illustrates some experimental results and in Section 6 the conclusions of the present work are drawn.

2 The collection of Border Records and the related metadata

The particular document collection is the archive of border records from the Civil Government of Girona, Spain.

All in all, it consists of 93 linear meters of printed and handwritten documents from 1940 to 1976. These records contain heterogeneous information about people arrested at the northern Spanish border. This information is of great importance in studies about issues related to the Spanish Civil War and World War II and is thus interesting from a cultural heritage preservation standpoint.

The collection is organized in bundles, each about 20-30 pages long containing any kind of documents related to a group of arrested persons. Most documents are official communications and reports; but also other documents can be present such as passports, letters from relatives or telegrams. The organization of the documents does not follow any specific structure, except for the first page of each bundle which is a form containing the names of persons appearing in the documents of the bundle. Some examples of such documents are shown in Fig. 1.

2.1 Objectives and challenges

Currently, several institutions have expressed their desire to index this collection. At the present moment, this work is done by archivists that process the documents manually and eventually take some notes. However, the collection is being scanned (at the moment about 34 000 pages ranging from years 1940 to 1944 are available in TIFF format). Our contribution in this sense consists in providing document analysis tools for facilitating this task, which results in a twofold scenario:

(i) **Document image analysis**: Provide tools for automatic metadata extraction from the documents (such as image processing, OCR, word spotting, etc.)

(ii) **Human-computer interaction**: Provide tools for facilitating the manipulation, visualization and annotation of the documents (such as sketch-based interfaces).

We work at the metadata space, where this metadata can be automatically, semi-automatically or manually obtained.

This article is mainly concerned with part (ii) and describes some aspects of the sketch-based interface aimed to annotate the document collection. Other components of the system have been reported in [10] and [11].

2.2 Metadata space

With the purpose of obtaining a representation of the knowledge stored in the collection we provide the experts with the means for populating the knowledge base in an efficient way using a sketch-based interface. We have identified a set of entities that appear in the documents:

- Persons (linked to a record, with associated nationality, age, names of relatives)
- Places (referring to a person's location)
- Events (with associated date, place and reason).

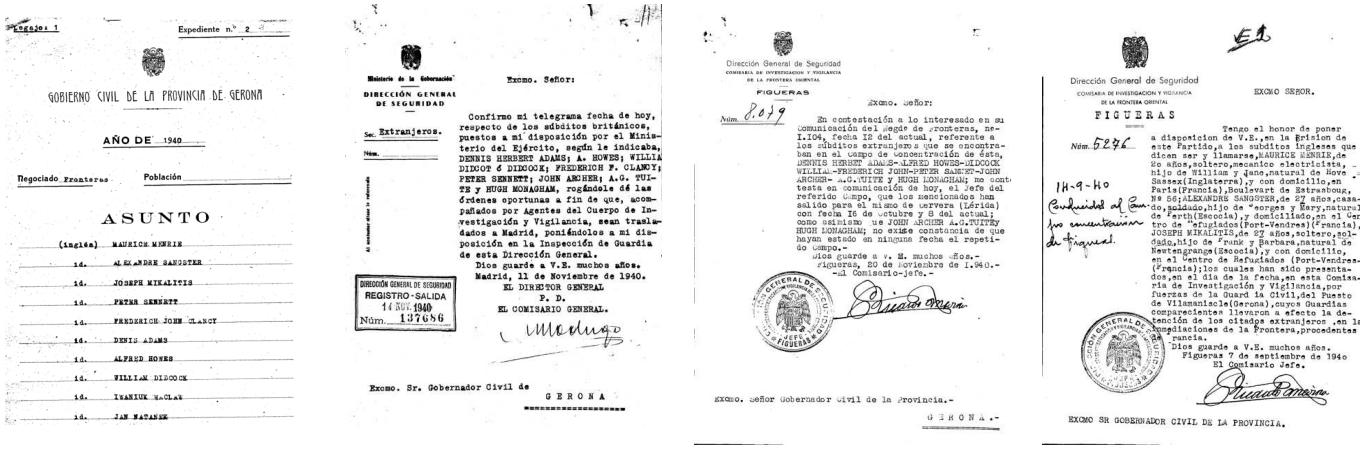
In addition to the annotation of the basic symbols, the user may establish relations between keywords and persons, places or events if the keyword carries information related to them. We may extract also relations between persons, among persons and events and relations between places and events.

The designed sketch-based interface should provide means for: (i) browsing the collection, (ii) selecting and annotating document contents, such as words (person names, keywords) or symbols, and establishing relationships between them, (iii) adding notes (typed or handwritten) to the selected contents or to the whole document, and (iv) launching automated processes.

The details of the proposed sketch-based interface are given in the following sections. First, we review the previous work that has been done in terms of automatic metadata extraction via word spotting and we explain how it can be complemented by the proposed sketch-based interface. Then, we will introduce the complementary sketch-based interface for semi-automatic annotation.

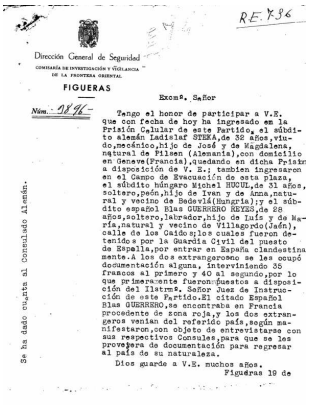
3 Word spotting for knowledge extraction

Informally speaking, the metadata associated to document images is a semantic graph structure whose nodes represent key concepts and the arcs represent semantic relations. The knowledge that is stored in the nodes is related to keywords in the document images. Keywords refer to names, dates, places, etc. Afterwards, the human intervention allows to complete the knowledge with semantic annotations. The process of *word spotting* aims to locate query keywords in document images. Two strategies exist. First, transcribing the document using OCR and a subsequent inexact string search process to locate the keyword. But in some documents, commercial OCR engines are not able to succeed. This is due to heterogeneous layouts (e.g. text combined with graphics artifacts), or due to document aging that results in degraded and noisy images. In such cases, some authors have developed word spotting techniques to search keywords at image level. Thus, keywords are modelled with shape signatures in terms of image features. The detection of the keyword in a document image is done by a cross-correlation approach between the prototype signature and signatures extracted from the target document image. A number of contributions exist in the literature on word spotting methods both for typewritten documents [8, 12], and handwritten documents [15, 18].



(a) (b) (c) (d)

Figure 1. Sample images of the document collection. (a) An example of record index. (b),(c) and (d) are official communications.



(a)

```
Direccin Generdi de Seguridad
COMISARA 11E INVESTIGACIN Y
VIGLANCiA
OE LA FRONTERA ORIENTAL
FIGUERAS Excma, Sefior
Nm...di.....Tengo
e]. honor de participar a
V.E. que con fecha de hoy ha
ingresado em la Prisi6n Celular
de este Partldo, el sbdito
alemn Ladlslaf *STEKA,d.e 32
afios,viudo,meonico,hijo de Jos y
de Migdalena, nqtural de Plisen
(Alemania) ,con domicilio en
Geneve(Francia) ,quedando en
dicha Pristr a disposici6n de
V. E.; tambien ingresaron en el
Campo do Evacuacin de esta plaza,
el sbdito hngaro Mjchol HUCUL,de
31 afios, soltero,po6n,hijo de
Ivan y de Anna,natural y vecino
de Bedevl(Hungria);y el sbdito
...
```

(b)

Figure 2. Errors when an off-the-shelf OCR is applied due to noise and heterogeneous layouts. (a) Original Image and (b) OCR transcription.

In the sketch-based annotation framework described in this paper, we perform a word spotting approach as the action associated to a sequence of user gestures. Thus, the user is able to select a keyword by cropping it in a document image. Afterwards, the user is able to spot it into the rest of the documents. A usual case is when the user selects a name in the cover page of a bundle, and the system looks for it in the other pages. We apply a combined strategy, namely word spotting at the text level and the image level. During the digitization process, all the documents of the corpus were OCRed using an off-the-shelf engine. Due to complex layouts and the presence of noise, the result of the OCR is very inaccurate as can be seen in Fig. 2. Thus, keywords are first searched in an alphanumerical mode by an inexact substring matching approach. The Levehnstein distance is used to model the character edit errors. Afterwards, in addition to overcome the text-level errors, a second word spotting approach has been developed at image-level. A compact shape signature inspired in the well-known shape context descriptor [2] is used to represent keyword images. Skeleton points are taken as feature points. Any given shape context h_i of a feature point p_i, it can be converted to a bit vector representation b_i by binarizing h_i by using a threshold T experimentally set. This bit vector of a shape context is called the *codeword* of the point p_i. Since our shape contexts are organized in 24 zones, distributed in three concentric rings, each codeword is a 24-bit number (see Fig. 3). For the sake of compactness, the codeword is split in three codes of 8 bits each $b_i = (\beta_i^1, \beta_i^2, \beta_i^3)$. These codes are used as hashing keys to index in a look up table that generate votes in a set of bins defined over the image. Those bins having a high number of votes are zones likely to contain the queried keyword. The reader is referred to [11] for a further description of this method.

4 Sketching interface for knowledge capture from experts

The development of sketch-based interfaces has received a growing interest in the last two decades due to the advent of new devices that allow a better human-computer interaction (HCI). Document Analysis Systems benefit from such interfaces in the interactive editing of documents. Thus, the

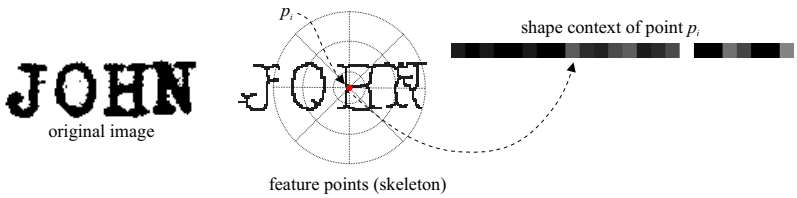


Figure 3. Representation by a Shape Context.

user is able to edit the document components by freehand sketches that the computer interprets and converts to actions performed on the document contents. One of the key advantages of a sketch-based interface over a keyboard-mouse one is the fact that the user feels more confident due to the similarities with a pen and paper. Two main key properties of sketch-based interfaces are the response time to user actions and how well the system is able to cope with ambiguity due to inaccurate strokes.

The aim of the sketch-based interface presented in this paper is to provide an application to extract knowledge from the collection of documents described in section 2. Each of the symbols defined in the alphabet is recognized using a syntactic approach and the corresponding associated action is executed. Depending on the associated action, we may divide the alphabet of symbols into four categories: navigation/browsing, selection or knowledge capturing, relational and handwritten annotations. Let us further describe this alphabet.

Navigation/Browsing symbols allow the user to navigate or explore the collection, going forward or backward, zooming in and out. Figure 4(a) shows the set of navigation gestures.

The symbols forming the *Knowledge extraction* set allow the user to select an image zone and to invoke an action such as OCR transcription, word spotting, annotating or providing a meaning for a part of the image. The set of symbols composing this category are shown in Fig. 4(b).

Relational symbols allow the user to link two different selected zones of the image. The user can then indicate the kind of relation that exists between the two symbols. See Fig. 4(c) for more details on the set of symbols that form this category.

Finally, the last set of symbols consists of those symbols that have not been classified in any of the previous sets. This input is taken as a handwritten annotation and is sent to a recognition engine trying to transcribe it. Otherwise, it is just stored as a collection of graphical strokes.

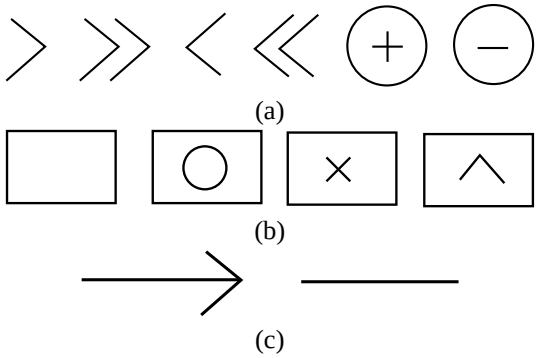


Figure 4. Set of gestures composing the different alphabets of the system: (a) Browsing elements, (b) Knowledge capturing and (c) Relational symbols.

4.1 Application Outline

The sketch-based annotation framework presented in this paper consists of four components. The main module browses the corpus documents and is able to capture the gestures done by the expert. The second component that recognizes the gesture according to the above described symbol classes. The third component, the interpretation part, executes the corresponding action to the current data given the corresponding gesture category. Finally, the database component. The interface is connected to the document database presented in section 2. This is an important component of the system because user actions directly affect it: browsing the document collection, annotating, transcribing, etc. It is divided into two inter-related databases. While the first contains the set of documents to be annotated, the second contains the metadata extracted from those documents. The information contained in this database has been presented in section 2.2. Figure 5 shows the architecture of the presented sketch-based framework. Let us further describe the different components.

First our system captures strokes, i.e. basic elements consisting of the set of points captured by the system between a pen down and a pen up event. Before associating those strokes with the corresponding gesture, the system needs to preprocess them. The preprocessing step is dependent on the recognition technique. In the case of a syntactic or structural method, primitives must be extracted from the symbols. In the case of a statistical method, the system must normalize the set of strokes to produce a uniform

representation to better compare with the models learned by the system at a previous stage.

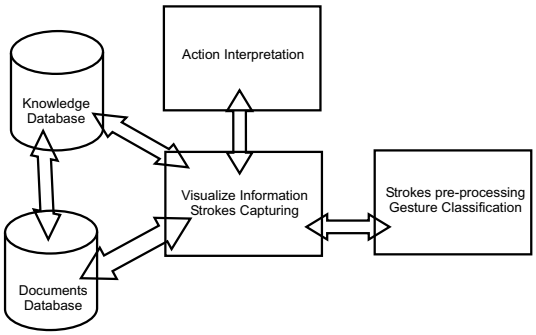


Figure 5. Schema of our Sketching Interface.

The recognition of gestures is done by means of the combination of two techniques. On one hand, there is a set of symbols that are recognized using a statistical technique, in particular a support vector machine (SVM) [5]. On the other hand, symbols defined as the combination of two or more basic symbols are detected with a syntactic approach. A syntactic approach consists of two main components, namely a grammar that describes the different symbols, and a parser that analyzes a set of strokes drawn by the user and assesses whether the symbol belongs to the language defined by the grammar, $L(G)$.

One of the key features of a sketch-based interface is that the user expects a quick response from the system to an action. To achieve this purpose, we need to recognize composed symbols in real time.

When a symbol cannot be recognized as belonging to one of the described categories (browsing, knowledge extraction and relational), the strokes are treated as handwritten annotations and a commercial on-line handwriting recognizer is used.

4.2 Recognizing Gestures and composed symbols

As mentioned in the previous section, two different techniques have been developed in the system for recognizing graphical gestures. The first one is based on a statistical approach. The approach uses scale invariant moments for representing the symbols and SVM to classify them, based on the approach we presented in [16]. Thus the system includes a prior training step to learn the different instances of the symbols.

The other approach is based on a syntactic methodology. Concerning the grammatical formalism, we have used an adjacency grammar [6] [14]. Productions in this kind of grammars are formed by a set of symbols and relations among them. Contrary to grammars for textual languages they do not expect a predefined order in the appearance of these symbols and for that reason they are useful in sketch-based interfaces. The grammar used in this work is presented in Fig. 6. Productions are presented as a set of primitives as circles, rectangles, etc. and relations among them as be inside, being horizontal, etc. The parsing algorithm selected to interpret the different symbols drawn by the user is based on a grid-directed algorithm. This fact allows the parser to restrict the search space to a grid-based neighbourhood in the process of interpreting valid compound symbols. For further details on the syntactic symbol recognition approach the reader is referred to [14].

OCR := {Rectangle, ce} IsInside(Rectangle,ce)
Spotting := {Rectangle, Cross} IsInside{Cross, Rectangle}
Cross := {Seg, Seg} Intersects(Seg, Seg)
Annotate := {Rectangle, Symbol2} IsInside(Rectangle, Symbol2}
Relation := {Rectangle1, Seg, Rectangle2}
hasVertexInside(Rectangle1,Seg) && hasVertexInside(Rectangle2,Seg).
Zoomin := {ce, Cross} isInside(ce,Cross)
Zoomout := {ce,Seg} isHorizontal(Seg) && isInside(ce,Seg)

Figure 6. Grammar corresponding to the system's alphabet.

4.3 Recognizing handwritten annotations

The elements that cannot be classified into one of the defined symbol categories up to a certain confidence threshold are interpreted as handwritten annotations. At the present moment, we employ a commercial handwritten recognizer to obtain the text transcript. Both the text transcript and the original symbol are stored in the database.

5 Experimentation

Our work is still in progress. In this section we show qualitative examples of the usefulness of our framework for an efficient semi-automatic annotation. In particular we show the different possibilities to extract metadata according to the interpretation of the gestures alphabet. We do not report quantitative performance rates of the involved recognition methods, both namely word spotting, and symbol recognition, because these have been reported in previous work [11, 14, **?**]. We report four examples on different use cases applied on the corpus described in section 2. In the first the user associates semantic metadata (see section 2.2) to a document by knowledge extraction and relational symbols. The second example shows a spotting procedure using the technique presented in section 3. The third example shows the transcribed text arising from the OCR module over a selected area. Finally, the last example shows how the user may enrich the metadata with handwritten annotations.

Figure 7 shows a possible process followed by a user to extract metadata from a document. In this process, the user first selects a region of pixels and associates the type *person* to it, see Figs. 7(a) and (b). Afterwards, the user selects another region. Finally a relational symbol is drawn between the regions, that identifies the second region as the age of the person defined in the beginning. These steps are illustrated in Figs. 7(c) and (d).

Concerning the spotting techniques, Fig. 8(a) presents the results of the corresponding user gesture. In this case, the user selects a zone on the image where the name *ERIC BRYANT* appears in the first page of a bundle, like the one presented in Fig. 1(a) and the process returns a page where the word is likely to appear. In this case, we can observe that the returning page contains the word. This result could be used to augment the metadata of the document with the process presented before.

In Figure 8(b) an example of transcription is shown. The user has selected a zone on the image where printed text appears, then the system sends the cropped image of this zone to the OCR that returns the transcription on the right.

Finally, Fig. 8(c) shows how the user is able to augment the metadata of a document by adding a handwritten annotation. In this case, a zone of the image has been previously selected, either by the annotation process of Figs. 7(a) and (b) or by means a spotting technique. The user writes a handwritten annotation and the system recognizes it and inserts into the metadata database using the technique of section 4.3.

6 Conclusions

We have presented a semi-automatic annotation application for archival documents. The application consists of multiple tools driven by a sketch-based interface. Using this interface, an expert can efficiently browse the collection, extract metadata from the document images to populate a knowledge database, create new metadata including handwritten annotations and launch automatic extraction processes (OCR and word spotting).

With this wide range of possibilities, a significant part of the heterogeneous knowledge stored in the document collection can be now extracted and organized, and eventually be used for efficient retrieval purposes. For instance, another part of the project, not reported in this work, is a web-based interface for efficient visualization and retrieval of the collection. This part can obviously take advantage from the work presented in this article. Apart from this, the current system might also allow archivists to work in a digital environment, and computer scientists to create quality groundtruth for benchmarking recognition and retrieval methods.

It should be noted that the system is still under development since the document collection is large and challenging. As future work we consider not only to improve the efficiency of all the involved processes, but also to introduce new tools, such as word spotting in handwritten pages and document categorization.

Acknoledgements

Work partially supported by the Spanish projects TIN2006-15694-C02-02 and CONSOLIDER INGENIO 2010 (CSD2007-00018), and the *Subdirecció General d'Arxius de la Generalitat de Catalunya*.

References

[1] H. Baird. Digital libraries and document image analysis. In *International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 2–14, 2003.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(24):509–522, April 2002.

[3] F. L. Bourgeois, E. Trinh, B. Allier, V. Eglin, and H. Emptoz. Documents images analysis solutions for digital libraries. In *Workshop on Document Image Analysis for Libraries (DIAL)*, pages 2–24, 2004.

[4] H. Cao and V. Govindaraju. Vector model based indexing and retrieval of handwritten medical forms. In *ICDAR*, pages 88–92, 2007.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[6] J. Jorge and E. Glinert. Online parsing of visual languages using adjacency grammars. In *Proceedings of the 11th International IEEE Symposium on Visual Languages*, pages 250–257, 1995.

[7] L. Kara and T. Stahovich. Hierarchical parsing and recognition of hand-sketched diagrams. *Proceedings of the 17th annual ACM symposium on User interface software and technology*, pages 13–22, 2004.

[8] S. Kuo and O. Agazzi. Keyword spotting in poorly printed documents using pseudo 2-d hidden markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):842–848, 1994.

[9] J. Landay and B. Myers. Sketching interfaces: Toward more human interface design. *IEEE Computer*, 34(3):56–64, March 2001.

[10] J. Lladós, P. Pratim-Roy, J. Rodríguez, and G. Sánchez. Word spotting in archive documents using shape contexts. In *3rd Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA2007)*, pages 290–297, 2007.

[11] J. Lladós and G. Sánchez. Indexing historical documents by word shape signatures. In *9th International Conference on Document Analysis and Recognition*, pages 362–366, 2007.

[12] S. Lu and C. Tan. Retrieval of machine-printed latin documents through word shape coding. *Pattern Recognition*, 41(5):1816–1826, 2008.

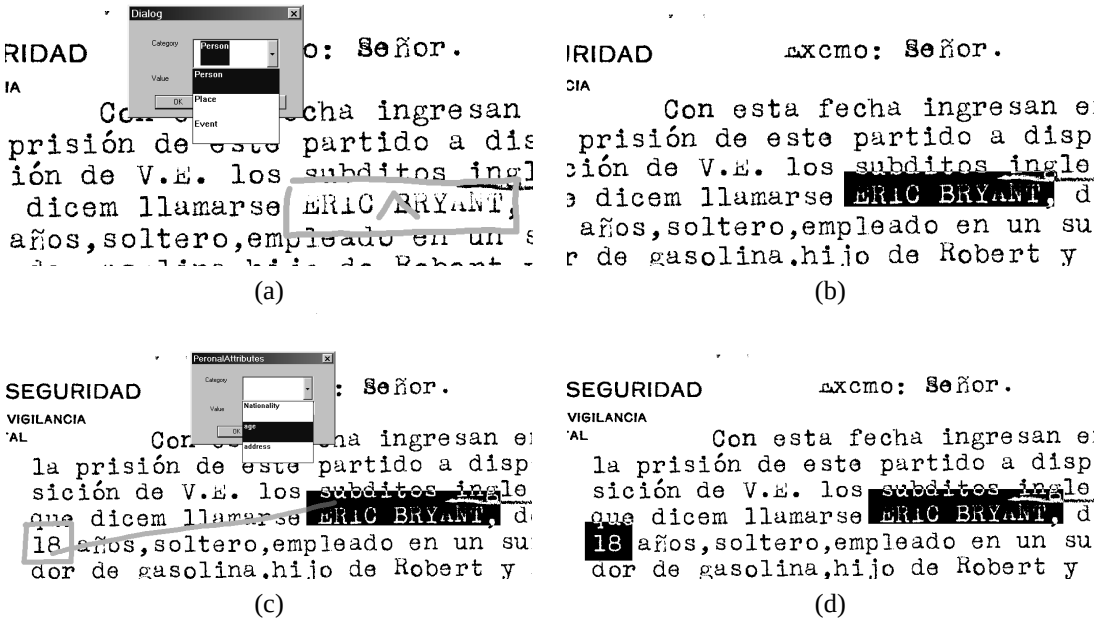


Figure 7. Annotation process: (a) and (b) Definition of a region indicating a person and (c) and (d) Creation of an attribute relation.

[13] R. Manmatha, C. Han, and E. M. Riseman. Word spotting: A new approach to indexing handwriting. In *Proc. of the 1996 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 631–637, 1996.

[14] J. Mas, G. Sanchez, and J. Llados. An incremental parser to recognize diagram symbols and gestures represented by adjacency grammars. In J. L. W. Liu, editor, *Graphics Recognition: Ten Year Review and Perspectives*, volume 3926 of *Lecture Notes in Computer Science*, pages 252–263. Springer-Verlag, 2006. DAG.

[15] T. Rath and R. Manmatha. Word image matching using dynamic time warping. In *Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 521–527, June 2003. Madison, WI.

[16] J. A. Rodríguez, G. Sánchez, and J. Lladós. A pen-based interface for real-time document correction. In *9th International Conference on Document Analysis and Recognition (ICDAR2007)*, volume 2, pages 939–943, September 2007.

[17] K. P. Sankar, V. Ambati, L. Hari, and C. Jawahar. Digitizing a million books: Challenges for document analysis. In *Workshop on Document Analysis Systems (DAS)*, volume 3872 of *Lecture Notes in Computer Science (LNCS)*, pages 425–436, 2006.

[18] C. Tomai, B. Zhang, and V. Govindaraju. Transcript mapping for historic handwritten document images. In *Proc. of 8th International Workshop on Frontiers in Handwriting Recognition*, pages 413–418, Aug 2002. Ontario, Canada.

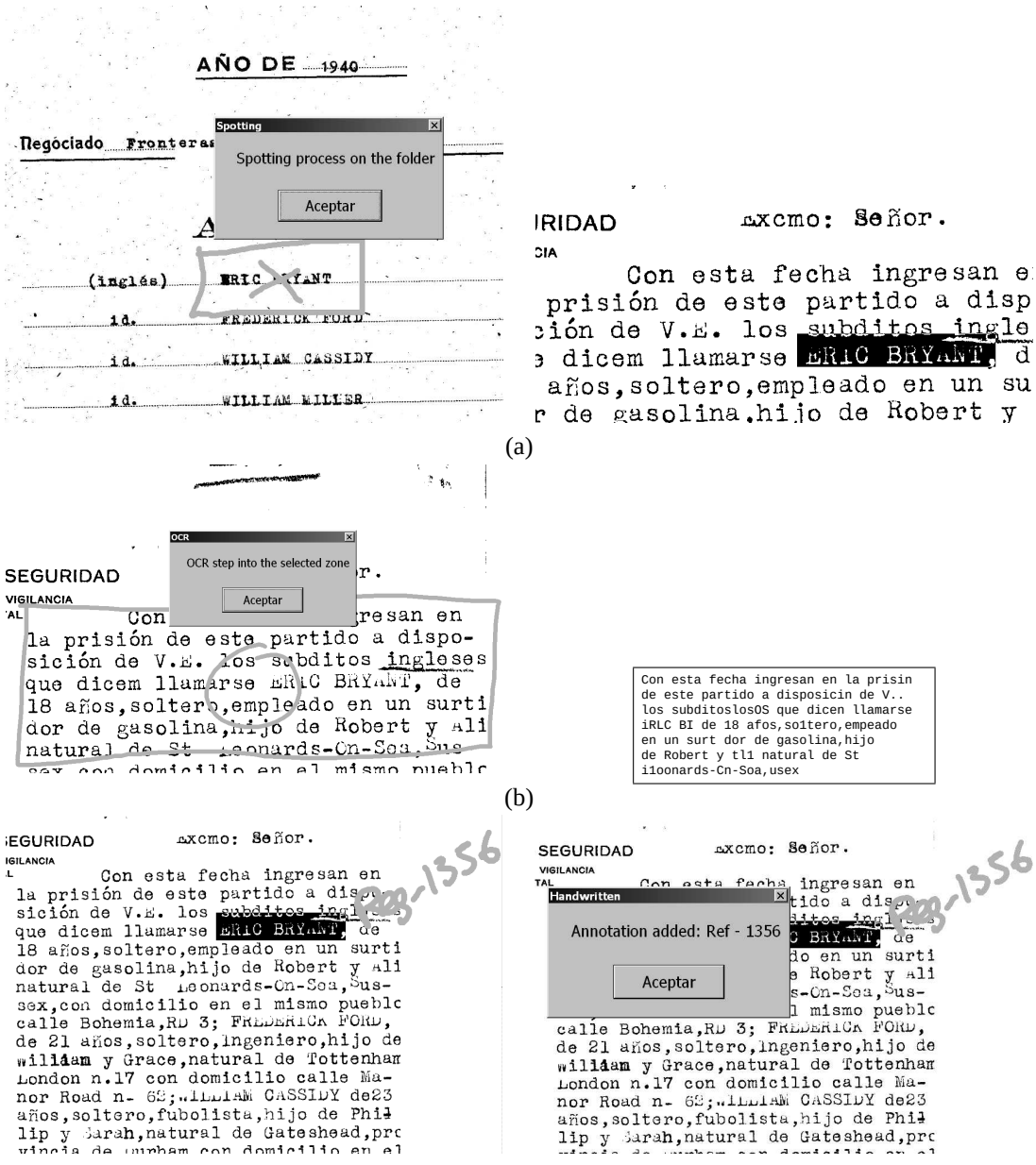


Figure 8. Screen shots showing the application's response to different gestures: (a) Spotting Gesture, (b) OCR Gesture and (c) Handwritten annotation.