Character Recognition Approach Based on Ontology

Hacene Belhadef\textsuperscript{1}, Aicha Eutamene\textsuperscript{1}, Mohamed Khireddine Kholadi\textsuperscript{1}

\textsuperscript{1}Mentouri University of Constantine, Department of Computer Sciences, Road of Ain El Bey, Constantine 25017, Algeria

\{hacene\_belhadef, aicha.eutamene, kholladi\}@umc.edu.dz

ABSTRACT

In this paper, we present a general description of a new character recognition process based on ontology, semantically annotated by a domain expert. Such process is based essentially on a matching step between two ontologies, the first represents a domain ontology, containing the typographical description of different characters represent an alphabet of a well definite language (Latin for example), the second ontology, describes the document in question in the form of concepts where each concept represents a grapheme located in a defined order in the document. In this work we consider that there is a robust extraction module of primitives that will serve us to instantiate our ontology, by the different graphemes and their spatial relationships that exist in the processed document.

KEYWORDS

character recognition, grapheme, matching, ontology, semantic annotation, typographical features.

1 INTRODUCTION

The general purpose of character recognition either printed or handwritten is transformed into a readable representation and exploited by a machine. The recognition process is not always easy as long as the contents of the documents may have multiple representations. In the case of printed documents, size, style (Bold, Italic... etc), and melting characters and other factors play a crucial role in such a process. As for handwritten documents, the conditions of safeguard are often not adequate. Today a large number of books and manuscripts are preserved in museums and archives and are at risk of disappearing due to several factors such as moisture, acidity ..., which requires scanning these documents in order to preserve the heritage and exploit a more efficient manner. The digitization of documents is the most effective and fast for this problem, it is to convert a document in paper form into a digital image. Transcription is another solution, but it is less used and limited to manuscripts documents not-long.

The result image of such operation of digitization is used as raw material in the recognition process, to decorticate the content and extract the necessary primitives for the identification and characters recognition, also the entire contents of the document, to use it in a lot of area such as the restoration of national heritage or world, classification, indexing and archiving.

Whatever the rich content of the documents, but this wealth is still insufficient to help the process of character recognition. In this article we are aware that one step annotation document is required to add information further helping this process to accomplish its task.

The annotation of an image through the construction of ontology constitutes the main tool for associating semantics to an image and allows the use of research
methods more powerful and able to answer complex queries. The association between data and ontology then allows the software agents to take advantage of the knowledge represented in ontology to better use the images. There are several annotation types; Figure 1 shows our choice for semantic type annotation on an ontology created after a step of segmentation and feature extraction.

This article is organized as follows. In First, we present an overview on the two concepts strongly linked to our process, as the ontology and annotation. In the paragraphs that follow, we present a description of the stages of our process scanning until the last step is the post processing showing our contribution to enrich this process.

2 ONTOLOGY

The concept of ontology is a concept that is not always easy to characterize. Indeed, it is used in different contexts: philosophy, linguistics, intelligence (AI), and each one's have its particular definition. The most commonly accepted definition is that given by T. Gruber in 1993: "a specification of a conceptualization to help programs and humans to share knowledge " [1]. More specifically: "... the Ontology is an explicit formal specification of a shared conceptualization. » According to Gruber," a conceptualization is the result of modeling phenomena in the domain of interest. This model identifies concepts and relations describing these phenomena. "Explicit" means that the concepts and their relationships are typed, and constraints on the use of these types are clearly explained. "Formal" refers to the fact that ontology should be understandable by the machine. "Shared" reflects the idea that ontology should capture the knowledge commonly accepted by consensus of all community stakeholders in the field [2].

The preceding definition leads to a set of definitions that can be used as a basis for algebraic formulation of the term Ontology and its components [3]:

Definition 1. A term is a triple $\tau = [\eta, \delta, A]$, $\tau \in T$, where $\eta$ is a string of characters containing the name of the term, $\delta$ is a string of characters containing its definition and $A$ is a set of attribute domains $A_1, A_2, ..., A_n$, each associated to a value set $\mathbb{V}$.

Definition 2. A relation $\phi : T \rightarrow T$, $\phi \in \Phi$, is a function from $T$ to $T$ such that for every term $\tau_1 \in T$, there is a term $\tau_1 = \phi(\tau_1), \tau_2 \in T$.

Definition 3. A semantic relation $\sigma$ between two terms is a relation that belongs to the set of semantic relations $\Sigma = \{\text{Hypernymy, Hyponymy (is-a), Mereonomy (part-of), Synonymy}\}$, $\Sigma \subset \Phi$.

Definition 4. A spatial relation $\rho$ between two terms is a relation that belongs to the set of spatial relations $P = \{\text{adjacency, spatial containment, proximity, connectedness}\}$, $P \subset \Phi$.

Definition 5. An ontology is a pair $\Theta = [T, \Phi]$ where $T = \{\tau_1, \tau_2, ..., \tau_n\}$ is a set of terms, and $\Phi = \{\phi_1, \phi_2, ..., \phi_n\}$ and $\exists \phi_i \in (\Sigma \cup K)$.

3 CLASSICAL CHARACTER RECOGNITION PROCESS

The main steps of a recognition process are [4]:
- The acquisition enables the conversion of paper document in the form of a digital image (bitmap). This step is important
because it is concerned with preparing documents to be seized, the choice and parameterization of hardware input (scanner), and the format for storing images.

- Pre-processing whose role is to prepare the document image processing. Pre-processing operations are related to the recovery of the image, remove noise and redundant information, and finally the selection of appropriate treatment areas.

- Recognition of the content that often leads to the text recognition and extraction of logical structure. These treatments usually accompanied by preparatory operations of block segmentation and classification of media (graphics, tables, pictures, etc...).

- Post-processing or correction of recognition results, to validate the digitization process. This can be done either automatically by the use of dictionaries and linguistic methods of correction, or manually through dedicated interfaces.

4 CHARACTER RECOGNITION APPROACHES

In this section we present four different approaches in this area that are most used.

4.1 Bayesian approach

The Bayesian approach is based essentially on the properties of the typographic character. It consists in choosing among a set of characters, one for which the following primitive has the highest probability with respect to characters previously learned.

4.2 Structural Approach

Like the Bayesian approach, the structural approach is also based on the physical structure of characters. Indeed, it is to describe the nature of relations linking different topological primitives (a loop, arc…). These relationships can be primitive relative position compared to another type of features (vertical or horizontal), the size of a primitive compared to another. There are variants of the structural approach are methods based on different principles. The most commonly used are:

4.2.1 Syntactic Methods

These methods use the notion of language for recognition. Each character is described or represented by a phrase in a language where the vocabulary consists of primitives. Thus, recognition of a character is whether the sentence of the representative character may be generated by the grammar.

4.2.2 The Method of Graphs

This method is based on graph theory. It is to build graphs, where nodes represent the primitives and arcs, the relationships between these primitives. In the learning phase, graphs representing characters of reference are established. This gives for each character reference or model, which includes a graph representing the different primitives describing (node) and all relations between them (arc).

4.2.3 Stochastic Approach

A stochastic process implements specific probabilistic models in order to manage uncertainty and lack of information that plague the characters to recognize. Among the methods, we find the method of Freeman code, hidden Markov models for modeling of characters (strings of primitives) ... etc.
4.3 Neuronal Approach

Connectionist or neuronal approach is based on the behavior and the human brain, to achieve entirely new models. His philosophy is reflected in the design of systems, called neural networks, based on a strong interconnection of a large number of elementary processors or pseudo-neurons. The connections between pseudo-neurons (variable weight) and contain all knowledge.

5 SEMANTIC ANNOTATION

The annotation is used to associate to documents a rich description (metadata) additional, complementary, describing their content, based on consensual vocabularies and independent of applications. It may be as text, structured, or semantic. it fills the poverty in the semantic level provided by a document.

There are many semantic annotation tools that can be distinguished by the nature of information resources annotated (text, image, video, etc...), the user annotation (Automatic, semi-automatic or manual) and the ontology reference used [5]. There are simple tools for annotation, like Flickr and ACDSee [6] Allowing the addition of descriptions textual content. Other tools, like [7] and [8] propose a more complex with a graph concepts and spatial ontology in RDF [9]. These tools allow annotations on the image, its areas, and spatial relationships between them ("A is the area right of the region B)... The annotation on the content of manuscripts is to describe what's in an image, for example, objects and characters that appear on the image, relationships between character forming words and sentences. However the notation on the context is used for describing the image using these contextual elements. The date or the age and the conditions under which, the document was written. In the field of digital cameras, a tool that is widely used for annotating contextual image archive EXIF (Exchangeable Image File) When taking a photo, most digital cameras add to the header image at EXIF description. This description has metadata on the photo (name, size, date) on the camera (model, maximum resolution) and the settings used to shoot (flash intensity, the aperture ) [10].

As for our annotation module (Figure 1), the reader may refer to section 4.5 for more information.

6 DESCRIPTION OF OUR SYSTEM

In this section we describe the different steps of our proposed process, including the steps of annotation and mapping that distinguish between such process and another classic, located in the literature. Figure 1 shows these steps starting with digital acquisition and arriving at post-treatment through segmentation, extraction of features and other steps.

6.1 Acquisition and Digitizing

The acquisition allows the conversion of paper document in the form of a digital image. This step is important because she cares about preparing documents to be seized, the choice and configuration of hardware input (scanner) and that the format for storing images [11]. Systems acquisition of the most common, are essentially linear scanners or cameras (CCD). The resolution of the scanned image influence the steps a subsequent recognition system. It is commonly accepted that the optimal resolution of an image depends on the thickness of the line of writing. Thus, for subsequent treatments can be applied correctly, it must the trait of writing has a minimum thickness of 3 pixels, an image resolution of between 100
and 300 points per inch (8 to 12 pixels / mm).
Current scanners have a maximum resolution is 300 dpi\(^1\). They allow image capture of text or binary or grayscale gray is the same color [12]. Convert a paper document into electronic document is the purpose of document recognition systems. This task is a necessity for access to a wider audience and offer new services such as preservation of ancient documents, consultation documents rare, Rapid duplication and economic structures, research information more efficient and finally the possibility of share knowledge, [13], [14] ... etc.

\(^1\) dpi: dots per inch
6.2 Pre-processing

The pre-processing included all preparatory operations image processing; its role is to prepare the image document for further process steps. It includes all functions performed by the segmentation and feature extraction to produce a version cleansed of the original image so that it can be used directly and effectively. This step is not specific to a given type of writing but it is part of any system pattern recognition to improve the quality of information for process steps that follow. So the pre-processing’s are used to eliminate the defects associated with scanned image; these defects can be of two types: those related to the channel scan (tilt brightness, noise ...) or those related to quality intrinsic to the document (the wet spots, appearance of back, holes ...) [15].

6.3 Segmentation

Segmentation is an essential step in the recognition process. It conditions strongly its robustness and determines its approach to a priori recognition. There are therefore two approaches: the comprehensive and analytical approach. To avoid segmentation problem of drawing a word into letters, the overall approach does not use segmentation and therefore consider the layout of the word in its entirety as to recognize the shape. For cons, the analytical approach based on segmentation and seeks to isolate and then identify meaningful units of the path of a word priori to match letters that compose it. Therefore good segmentation allows the present system to recognize characters in good conditions, cons improper segmentation, in turn, will lead falling rate of recognition [16] [17]. The segmentation process is generally subdivided into two main stages, the first step is to horizontal segmentation is to locate the entity to recognize this entity may be a paragraph, a line, word or group of characters, the second being cons vertical segmentation, it is to locate items that comprise the entity we want to recognize these elements can be words or characters.

6.4 Extraction of Features

The extraction of feature is a crucial and very important in the recognition process, because subsequent treatments will no longer manipulate the original image but the results provided by this step. Its role locates the letters in the word and spatial relationships between different lexical units indicating their limitations left and right. These characteristics are generally classified into two families: the structural characteristics (loops, concavities, overruns, extreme points of the route, intersections, etc ...) and the statistical characteristics that derived measures of spatial distribution of pixels (zoning, moment invariants, Fourier descriptors, density points, contours, etc.).

The feature extraction of characters may carry on:

---

2 Resource Description Framework
6.5 Creation of ontologies

This step is to create two ontologies: domain ontology and document ontology.

**Document ontology**: This ontology represents the contents of the document; it is created after an operation of segmentation and extraction of primitives (grapheme), representing typographic forms of character of the processed document; where each concept of ontology represents a grapheme located in a defined order in the document.

Example of an excerpt from document ontology represents the word: “LIRE”

![Figure 1. Representation of the word “LIRE”](image)

The relationship between the grapheme number 1 (tall trunk) and 2 (bar), is clearly depicted in figure 3, and the corresponding code of this relationship is shown below.

**Domain ontology**: This ontology contains typographical features of character of the Latin alphabet, which are represented in the figure below. Figure 3, presents this ontology created under protégé editor [18]

![Figure 3. Typographical features (grapheme)](image)

6.6 Semantic Annotation

Before performing the step of annotation, one must first create the document...
ontology. This last, represents the different characteristics of contents processed document. The representation of documents has evolved classical representation is the representation of gross document to the structured representation. Representation Structured documents are usually described using the XML (Extensible Markup Language) are the presentation of XML schema or DTD (Document Type Definition). Contrary to the conventional representation of a document that reflects only the text content and information is presented in a flat without any treatment, the structural presentation focuses on the presentation of structured content, and all information is associated with a label, be logical or physical. The physical structure describes the organization of the document (Lines, Paragraphs, Images...) thus that the format of information (Center, Left, Right...) is the description of the appearance of document to the user. However the logical structure, presents the significance of this information (Title, Name_author, Date...). [19] In our case the annotation is performed on the ontology document written in OWL (Web Ontology Language) the purpose of this annotation is the enrichment of the ontology by semantic information that are not on the document and can query the ontology by and the inference mechanism to facilitate and accelerate the recognition process.

6.7 Matching

The process of matching is an algorithm that takes as input parameters: the schema Target (ontology document) and the destination schema (reference ontology). This process determines the values semantic similarities between elements and attributes patterns. The objective here is to find correspondences nodes of the source and target ontology, that’s why the similarity measures should be used to estimate the similarity between the paired elements and the identification and recognition of characters represented by these nodes. This process can be used in the stage of recognition or in the stage of post-processing.

6.8 Classification and Recognition

Character recognition can decide on the identity of a character from a learning of its form. This step requires a prior step of parameterization of the form, defining data, measurements, or visual cues that underlie the recognition method. Depending on the nature of this information; there are several categories of methods: syntax (described by a grammar), structural (described by a graph) or statistical (descriptions by partitioning the space). These have by far the greatest interest with methods based on neural networks or stochastic model [11]. Recently several authors seek to extend recognition to the manuscripts recognition of words, sentences and entire texts. In [20] the author proposed an approach for recognition of characters and offers its adaptation for the recognition of handwriting, using a system based on a MRF modeling and programming dynamic 2D. In [21] the author uses a General Markov bi-dimensional approach, for document recognition manuscripts (AMBER)-based programming Dynamic and inspired by the method of Knight. This approach has allowed the build script mail [22]. Proposes a system of recognition of handwritten words Arab isolated, belonging to a lexicon, based on a hybrid system based on Neural Networks and Models Hidden Markov (HMM) and segmented into graphemes [15].
6.9 Post-processing

A document consists of a set of characters. Therefore, recognition of these characters implies recognition of the full text. But it often happens, for multiple reasons, the system is wrong or does not identify certain character, which renders some words invalid. The main goal of this phase is to improve the recognition result by making corrections spelling or using morphological tools references such as dictionaries, thesauri, ontology Contextual, etc.... The identification of lexical units in the document does not only concern primary system of recognition. The majority of recognition methods are limited to the analysis of physical structure and recognition logic remains to develop, for which we propose in our system add other steps, such as checking and syntax semantic content of documents, by defining a grammar or a contextual ontology.

In the field of heritage restoration, especially the old documents that have been preserved in non-appropriate conditions, these documents may experience problems such as lack of significant portions or deleted words or paragraphs that constitute the document. One solution proposed to solve this problem is the use of ontologies by verifying the semantic of the lexical and syntactic token.

7 CONCLUSION

The field of recognition of handwritten documents is rather quite recent, the work is growing and several issues remain open, although an abundance of research about this area. In this paper we have presented a new recognition process or rather new approach based on the creation of an ontology representing the contents of an handwritten document and a step of semantic annotation which tries to enrich this last ontology, then a step of matching with a domain ontology for character identification.

We also plan to create a global ontology that brings together all the Latin alphabet, and define the different spatial relationships between all typographical forms (grapheme) and defining appropriate similarity measures for the process of matching, and we plan in the step of post-processing, to use of an external resource such as WordNet\(^3\) for syntactical and grammatical checking.

REFERENCES


\(^3\) WordNet : A lexical database of English: http://wordnet.princeton.edu/

9. The World Wide Web Consortium (W3C), http: \ \www.w3c.org


