Case-Based Cleaning of Text Images

Éric Astier[1], Hugo Iopeti[2], Jean Lieber[3],
Hugo Mathieu Steinbach[2], and Ludovic Yvoz[2,3]

[1] Persée, École Normale Supérieure de Lyon, F-69007 Lyon, France
[2] Université de Lorraine, master informatique
[3] LORIA, Université de Lorraine, CNRS, Inria, F-54000 Nancy, France

Abstract. Old documents suffer from the passage of time: the paper is becoming more yellow, the ink is fading and the handling of these documents can still cause them to be damaged (e.g. stains may appear). The online edition of such documents can have profit of a cleaning step, the aim of which is to improve its readability. Some image filtering systems exist that require proper parameters to perform a cleaning. This article presents Georges, a CBR system designed to predict these parameters, for images of texts written in French, with several variants based on approximation, interpolation and extrapolation, based respectively of similarity, betweenness, and analogical proportion relations. The images are characterized by a dirtiness index, with the assumption that two images with similar dirtiness indexes would require similar sets of parameters for being cleaned. This index is based on the detection of the occurrences of a frequent French word on the pages and of averaging these occurrences. Human and automatic evaluations show that the proposed approach (with its variants) provides high-quality results.

Keywords: case-based reasoning, approximation, interpolation, extrapolation, image processing, image cleaning

1 Introduction

The notion of similarity between a source case and a target problem is well-known to play a key role in case-based reasoning (CBR [14]). This involves the question of what similar means for a given CBR application and this is linked with the problem-solving task (i.e. to the nature of the relation between a problem and a solution). Consider, for example, problems represented by grayscale images. This problem can be represented by an $m \times n$ matrix of nonnegative numbers, so a naive way to assess the similarity between two problems would be to compute a classical norm-based distance between matrices such as $(A, B) \mapsto \sum_{i=1}^{m} \sum_{j=1}^{n} |B_{ij} - A_{ij}|$.

In this article, Georges, an application of CBR is presented in which problems are represented by such grayscale images, but similarity would be poorly modeled using such a norm-based distance function. For this application, a problem corresponds to such an image and a solution is given by a triple of values parametrizing a filter to be applied on the image for having it cleaned. More

precisely, this image is obtained by scanning a page containing some text. Two problems would be similar if their "dirtinesses" are assessed to be close, where the dirtiness of an image characterizes how it should be cleaned.

After some preliminaries (Section 2), related studies are presented (Section 3). Section 4 presents the application context and explains that one of the steps of the chain of treatments consists in choosing a parameter triple, which has been done manually so far. Several CBR approaches are proposed to generate triples of parameters (Section 5), thus making the semiautomatic cleaning step automatic. Section 6 evaluates and compares these approaches. Section 7 concludes, discusses the scope of this approach beyond the application, and points out some future work.

2 Preliminaries

This section presents some notions and notations in the domains of mathematics, CBR and image processing, useful for the remainder of the article.

2.1 Some mathematical notions and notations

Let $\mathbb{N}$ be the set of natural integers, $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$, $\mathbb{R}$ be the set of real numbers, $\mathbb{R}_+ = \{u \in \mathbb{R} \mid u \geq 0\}$, and $\mathbb{R}_+^* = \mathbb{R}_+ \setminus \{0\}$.

A *distance function* in a set $\mathcal{U}$ is a mapping $\texttt{dist} : \mathcal{U}^2 \to \mathbb{R}_+$ such that (for $u, v, w \in \mathcal{U}$): (1) $\texttt{dist}(u, v) = 0$ iff $u = v$, (2) $\texttt{dist}(u, v) = \texttt{dist}(v, u)$, and (3) $\texttt{dist}(u, w) \leq \texttt{dist}(u, v) + \texttt{dist}(v, w)$. For $n \in \mathbb{N}^*$, let $\mathcal{U} \subseteq \mathbb{R}^n$ and `dist1` and `dist2` be the two distance functions defined, for $(u, v) \in \mathcal{U}^2$ by

$$\texttt{dist1}(u, v) = \sum_{i=1}^{n} |v_i - u_i| \qquad \texttt{dist2}(u, v) = \sqrt{\sum_{i=1}^{n} (v_i - u_i)^2}$$

For $u \in \mathcal{U}$, $\|u\|_{\texttt{dist1}}$ denotes the value $\texttt{dist1}(u, \vec{0})$, where $\vec{0} = (0, 0, \ldots, 0) \in \mathbb{R}^n$.

The notion of betweenness can be used to apprehend the ternary relation "a is between b and c" where a, b and c are objects of the same set $\mathcal{U}$. In [1], this kind of relations is studied in the framework of the Gärdenfors conceptual spaces [5]. In particular, some postulates for such a relation are given and one of the ways to define a betweenness relation based on a distance function on $\mathcal{U}$ is introduced.[1] It is reformulated as follows. Let $[b, c]_{\texttt{dist}}$ be the set of $a \in \mathcal{U}$ such that $\texttt{dist}(b, c) = \texttt{dist}(b, a) + \texttt{dist}(a, c)$. For example, if $\mathcal{U} = \mathbb{R}^n$, $[b, c]_{\texttt{dist2}}$ is the segment line $[b, c]$ in the usual sense and $[b, c]_{\texttt{dist1}} = [b_1, c_1] \times [b_2, c_2] \times \ldots \times [b_n, c_n]$ where $[b_i, c_i] = [b_i, c_i]_{\texttt{dist1}}$ is the set of $a_i \in \mathbb{R}$ such that $\min(b_i, c_i) \leq a_i \leq \max(b_i, c_i)$.

[1] The authors of [1] also criticize this definition by the fact that two distances on $\mathcal{U}$ defining the same topology do not necessarily correspond to the same betweenness relation. However, this is the definition considered in this paper, because it has the advantage of simplicity.

Then the $\texttt{dist}$-betweenness relation is defined, for $a, b, c \in \mathcal{U}$ by a is between c and d if $a \in [b, c]_{\texttt{dist}}$.

An *analogical proportion* on a set $\mathcal{U}$ is a quaternary relation $\mathcal{AP}$ on $\mathcal{U}$ verifying some postulates (not detailed here: see, e.g., [13] for details). For $a, b, c, d \in \mathcal{U}$, $\mathcal{AP}(a, b, c, d)$ is usually written as $a : b :: c : d$. In this paper, the only analogical proportion that is considered is the *arithmetic analogical proportion*, defined on a subset $\mathcal{U}$ of $\mathbb{R}^n$ (for some $n \in \mathbb{N}^*$) as follows:

$$a : b :: c : d \text{ if for every } i \in \{1, 2, \ldots, n\},\ b_i - a_i = d_i - c_i$$

An *analogical equation* is an expression of the form $a : b :: c : y$ where $a, b, c \in \mathcal{U}$ and y is a symbol called the unknown. Solving this equation consists in finding all the bindings of y by values d such that $a : b :: c : d$. For the arithmetic analogical proportion, such an equation as at most 1 solution.

If $\mathcal{U}$ is a set of $m \times n$ matrices on real numbers, the above definitions (distance functions, betweenness relations, analogical proportions) also apply, the only difference being that the indexes in the definitions are $(i, j) \in \{1, 2, \ldots, m\} \times \{1, 2, \ldots n\}$ instead of $i \in \{1, 2, \ldots, n\}$.

2.2 CBR: notions, notations and assumptions

The way in which CBR is presented in this section is strongly biased by the work presented in this article. For instance, a case is given by any problem-solution pair (and this is not true for every CBR application). Therefore, it should not be used as a short introduction to CBR in general.

Let $\mathcal{P}$ and $\mathcal{S}$ be two sets. A *problem* (of the current application domain) is by definition an element of $\mathcal{P}$ and a *solution*, an element of $\mathcal{S}$. A *case* is a pair $(\mathtt{x}, \mathtt{y})$ where $\mathtt{x} \in \mathcal{P}$ and $\mathtt{y} \in \mathcal{S}$. The case $(\mathtt{x}, \mathtt{y})$ can be read as the statement "$\mathtt{y}$ is a solution of $\mathtt{x}$." However, given two cases $(\mathtt{x}, \mathtt{y}^1)$ and $(\mathtt{x}, \mathtt{y}^2)$ sharing the same problem, the two solutions $\mathtt{y}^1$ and $\mathtt{y}^2$ are not equal: it is assumed that there exists an expert assessment of the quality of a solution $\mathtt{y}$ to a problem $\mathtt{x}$ (from the worst solution to the best one), where the expert is a human who can solve problems of this domain. This assessment is not known by the CBR system, except for the case base, which is a finite set $\mathtt{CB}$ of cases $(\mathtt{x}^s, \mathtt{y}^s)$ such that the expert assesses that $\mathtt{y}^s$ is a good solution to $\mathtt{x}^s$. A *source case* is an element of $\mathtt{CB}$.

Let $\mathtt{x}^{\mathtt{tgt}} \in \mathcal{P}$ be a problem to be solved, called the *target problem*. The CBR solving session of $\mathtt{x}^{\mathtt{tgt}}$ consists of two steps. *Retrieval* aims at selecting $k \geq 1$ source cases based on some criteria relative to $\mathtt{x}^{\mathtt{tgt}}$ (e.g. similarity to $\mathtt{x}^{\mathtt{tgt}}$, but this is not the only possibility). *Adaptation* uses these retrieved cases in order to propose a solution $\mathtt{y}^{\mathtt{tgt}}$ to $\mathtt{x}^{\mathtt{tgt}}$.

2.3 About image processing

The pixels of the images considered in this article are associated with gray levels. A *grayscale value* is modeled by a real number in the $[0,1]$ interval.[2] The gray level 0 (resp. 1) corresponds to a black pixel (resp. to a white pixel). An image $\mathtt{x}$ is modeled by an $m \times n$ matrix of grayscale values: a pixel of $\mathtt{x}$ is a pair $(i,j) \in \{1,2,\ldots,m\} \times \{1,2,\ldots,n\}$ and the grayscale value associated with a pixel (i,j) is $\mathtt{x}_{ij}$.

The treatment of an image considered in this paper is parametrized by a triple $(b\ell, w\ell, \gamma) \in [0,1] \times [0,1] \times \mathbb{R}_+^*$ and is the composition of 3 treatments, as described below.

Let $b\ell \in [0,1]$. Applying the "black level" transformation on a grayscale image $\mathtt{x}$ consists of substituting darker pixels according to the threshold $b\ell$ by a black pixel:

$$\mathtt{applyBlackLevel}(b\ell, \mathtt{x}) = \mathtt{x}' \text{ with } \mathtt{x}'_{ij} = \begin{cases} \mathtt{x}_{ij} & \text{if } \mathtt{x}_{ij} > b\ell \\ 0 & \text{else} \end{cases}$$

The "white level transformation" is described similarly (for $w\ell \in [0,1]$):

$$\mathtt{applyWhiteLevel}(w\ell, \mathtt{x}) = \mathtt{x}' \text{ with } \mathtt{x}'_{ij} = \begin{cases} \mathtt{x}_{ij} & \text{if } \mathtt{x}_{ij} < w\ell \\ 1 & \text{else} \end{cases}$$

The "Gamma transformation" is defined below, for $\gamma \in \mathbb{R}_+^*$:

$$\mathtt{applyGamma}(\gamma, \mathtt{x}) = \mathtt{x}' \text{ with } \mathtt{x}'_{ij} = \mathtt{x}_{ij}^{\gamma}$$

The effect of this transformation is essentially applied on the darker pixels: if $\gamma < 1$ these pixels get lighter, if $\gamma > 1$, they become darker.

Finally, given a triple $\mathtt{y} = (b\ell, w\ell, \gamma)$, the `clean` treatment consists of applying in sequence these 3 transformations:

$$\mathtt{clean}(\mathtt{y}, \mathtt{x}) = \mathtt{applyGamma}(\gamma, \mathtt{applyWhiteLevel}(w\ell, \mathtt{applyBlackLevel}(b\ell, \mathtt{x})))$$

In practice, a triple $(b\ell, w\ell, \gamma)$ should satisfy $b\ell \leq w\ell$, otherwise, the "cleaned" image would be fully white.

A *binarization* task on a grayscale image $\mathtt{x}$ consists in turning all pixels black or white. One way to do this consists of choosing a value $b\ell$ and computing $\mathtt{clean}((b\ell, w\ell, \gamma), \mathtt{x})$ with $w\ell = b\ell$ (and γ chosen arbitrarily). It should be noted that the goal of the `clean` treatment in our application is *not* to make a binarization. Indeed, its objective is to improve the readability of the page by human readers, and intermediate gray levels are useful for this purpose. However, the related work that is, to the best of our knowledge, the closest to our work is a binarization, as the next section presents.

[2] In practice, grayscale values are integers ranging from 0 to 255, but, it has been chosen to normalize them in $[0,1]$: from a $g \in [0,1]$, the corresponding integer value is $\lfloor 255x + 0.5 \rfloor$.

3 Related Works

The area of image processing is broader than the one presented below. In particular, image interpretation aims at finding high level descriptions of images and some CBR approaches have been successfully applied to this task (see e.g. [11, 12]). CBR can also be used to help experts in image processing, by helping them to build a relevant sequence of tasks to be applied on an image [4].

In this article, the goal is to find transformations on images so that they are more readable for human beings and we have not found CBR approaches for this purpose. By contrast, there are many research on the binarization of images (see [17, 10, 2] for surveys).

The binarization approach of Ergina Kavallieratou and Hera Antonopoulou [7] appears to be close to the objective of our work and this, for two reasons. First, they have worked on text images of old documents (in their work, handwritten or typed documents in Greek, which differs from the typed documents in French). Second, the main step of their process transforms grayscale images into grayscale images giving a cleaned image that is not black and white (the binarization takes place after this main step). The algorithm of this main step describes a cleaning procedure that is described briefly as follows, given an input image $\mathtt{x}$. First, a histogram of the grayscale values that are below the average value of the pixels is computed (this histogram is a function associating to $g \in [0, 1]$ the number of (i, j) such that $\mathtt{x}_{ij} = g$). Then, this histogram is rescaled, so that minimum values get to 0 and maximum values get to 1 (the darkest pixels become black, the lightest pixels become white) and the image is changed accordingly. This process is repeated until the difference between two successive images gets lower than a given threshold.

A qualitative comparison of the result of the process presented in [7] and the result of the process presented in this article is given in Section 6.4.

4 Application Context

Persée is a support and research unit of *ENS de Lyon* and CNRS assisting research by ensuring enriched digitization, quality processing, open dissemination and long-term preservation of scientific documentary heritage written in French. The missions of Persée are structured around 3 axis:

- Enhancing the value of collections of scientific publications through the Persée portal (`https://www.persee.fr`);
- Design, production and dissemination of research corpora (`https://info.persee.fr/section/perseides`);
- Making data available in a triplestore (`https://data.persee.fr`).

Given a document to be edited (a book, a journal, proceedings of a conference, etc.), its digital publication by Persée follows a workflow of tasks that are described below in a partial and simplified way:

(T1) The document is digitized (either by Persée or by the client) and put in an appropriate image format, i.e. a sequence of bitmap images in grayscale.
(T2) An OCR is run on all images, which extracts its words and there respective positions on the image.
(T3) The image is cleaned in a semi-automatic way:
  (T3.1) A page of the document is chosen (typically not one of the first or last pages with the cover, the table of contents, etc.).
  (T3.2) The expert chooses a triple $\mathtt{y} = (b\ell, w\ell, \gamma)$ for the image $\mathtt{x}$ of this page, visualizes the result $\mathtt{clean}(\mathtt{y}, \mathtt{x})$ and makes changes to the triple if the result is not adequate.
  (T3.3) The filter `clean` is applied to all images in the document with the same triple $\mathtt{y}$ (assuming that the degradation of the images over time is the same on all pages of the document).
(T4) The next steps are not described here. It is sufficient to know for this article that some further verifications are made that may lead to go back to step (T3) (which is relevant, in particular, when a page of the document has been degraded significantly more than the other pages).

5 CBR approaches to case-based text image cleaning

This section presents several approaches for automatizing the cleaning step (step number (T3)) presented in the previous section and gathered in the CBR system Georges. Section 5.1 describes the representation of the cases and the constitution of the case base (Section 5.1). Then, the question of comparing images from the cleaning task viewpoint is addressed (Section 5.2): the images are characterized by a "dirtiness index" and the comparisons between the source images and the target image are based on this index. Section 5.3 describes the application of several CBR approaches to this problem. These approaches have been introduced in [8] and the current article also provides an opportunity to study them within a concrete application domain.

5.1 Representation of cases and constitution of the case base

The task Georges has to solve is to associate to an image $\mathtt{x}^{\mathtt{tgt}}$ a triple of parameters $\mathtt{y}^{\mathtt{tgt}}$, such that $\mathtt{clean}(\mathtt{y}^{\mathtt{tgt}}, \mathtt{x}^{\mathtt{tgt}})$ is a cleaned version of $\mathtt{x}^{\mathtt{tgt}}$. Therefore, in this application, a problem $\mathtt{x}$ is a grayscale image, a solution $\mathtt{y}$ is a triple of parameters $(b\ell, w\ell, \gamma)$, and a case is an image-triple pair $(\mathtt{x}, \mathtt{y})$.

The step (T3) has been semi-automatic during several years, the experts associating to an image $\mathtt{x}$ a triple $\mathtt{y}$. Since the pairs $(\mathtt{x}, \mathtt{y})$ have been retained, they constitute cases, thus the case base is constituted by such cases: only a part of them have been kept for the case base (this is detailed in Section 6.1).

5.2 Representing image dirtiness

The images considered in this work are images of French texts (and potential illustrations), so the notion of cleanness (and, dually, the notion of dirtiness) is

(a) Excerpt of an image `x`, with 3 framed "de".

(b) Average "de" on `x`.

(c) `idx(x)` (variant 1×1).

(d) `idx(x)` (variant 2×3).

Fig. 1: The dirtiness indexing of an image of French text.

related to the readability of this text: the more a text image requires efforts to a reader, the more dirty it is.

It is noticeable that the goal here is not to assess a dirtiness level: two images could be qualified equally dirty, but could be dirty in different ways (e.g. one being too dark and the other one too light). It should also be noted that this representation of image dirtiness is related to the whole image, since the treatment at this step is global, but taking, e.g. the average grayscale of the page is a bad idea: it is easy to imagine two images having the same average of grayscale, one being rather clean and full of text, the other one being very dark, with a gray background, and having only a few lines of text.

So, the idea has emerged to qualify the dirtiness of an image using samples of the image that can be found (with various levels of dirtiness) in most images. Since these are images of French texts, a first idea was to use the occurrences of "e", the most frequent letter in French. In order to have a slightly larger number of pixels, the most frequent French word was chosen instead: the word "de".[3] Therefore, given an image `x`, the index of dirtiness of `x`, denoted by `idx(x)`, is characterized by the average "de" in `x`. More precisely, the indexing process is illustrated in Figure 1 and it can be described by the following process (in practice, these steps are combined, but are easier to explain this way):

1. From the OCR treatment of the image `x`, the set of bounding boxes of occurrences of "de" in `x` is collected (which necessitates some recalibration) and then, these bounding boxes are normalized by scaling, so that each occurrence of "de" corresponds to an $m \times n$ matrix of grayscale values.
2. Then, the average "de" is computed by a arithmetic mean of these normalized bounding boxes, giving MGV, a matrix of grayscale values.
3. The last step has two variants:

[3] This word corresponds approximately to "of". Its average number of occurrences by page in the studied corpus is 16.2, with a standard deviation of 11.5. For one of the rare pages in the corpus containing no "de", the backup solution is to use the solution of another page of the same document. Documents in French that contain no "de" are very rare, though one of them is famous: it is *La disparition*, a novel of Georges Perec which is about 300 pages lipogram in "e" and thus, contains no "de". This is a paradoxical explanation of the choice of GEORGES, as the name of our system.

(1×1) In the first variant, $\texttt{idx}(\texttt{x})$ is simply the average of grayscale values of MGV, i.e. $\dfrac{1}{mn} \displaystyle\sum_{1 \leq i \leq m} \sum_{1 \leq j \leq n} MGV_{ij}$.

(2×3) In the second variant, the matrix MGV is split in 6 parts by splitting the ranges of the line number i into 2 parts and the ranges of the line number j into 3 parts and taking the average of the grayscale values in each of the 6 parts. In other terms, $\texttt{idx}(\texttt{x})$ is a 2×3 matrix defined by:

$$\texttt{idx}(\texttt{x})_{pq} = \frac{6}{mn} \sum_{\left\lfloor \frac{(p-1)m}{2} \right\rfloor + 1 \leq i \leq \left\lfloor \frac{pm}{2} \right\rfloor} \sum_{\left\lfloor \frac{(q-1)n}{3} \right\rfloor + 1 \leq j \leq \left\lfloor \frac{qn}{3} \right\rfloor} MGV_{ij}$$

for $p \in \{1, 2\}$ and $q \in \{1, 2, 3\}$, with $\lfloor r \rfloor$ the floor of r.[4]

In the approaches presented in the following, the only information used in the images $\texttt{x}$ is given by $\texttt{idx}(\texttt{x})$. For a target problem $\texttt{x}^{\texttt{tgt}}$, computing $\texttt{idx}(\texttt{x}^{\texttt{tgt}})$ is a pretreatment of the processes. For $\texttt{x}^s$ such that $(\texttt{x}^s, \texttt{y}^s) \in \texttt{CB}$, $\texttt{idx}(\texttt{x}^s)$ is computed offline and stored in a database. For any source case $(\texttt{x}^s, \texttt{y}^s)$, $\texttt{idx}(\texttt{x}^s)$ is denoted by $\texttt{i}^s$; $\texttt{idx}(\texttt{x}^{\texttt{tgt}})$ is denoted by $\texttt{i}^{\texttt{tgt}}$.

5.3 Proposed approaches

The approaches presented in this section are based on the retrieval of k source cases, respectively with $k = 1$, $k = 2$ and $k = 3$, and the adaptation of these cases to solve $\texttt{x}^{\texttt{tgt}}$. These approaches correspond to the following reasoning scheme. It is assumed that $\mathcal{R}_{\texttt{pb}}$ (resp. $\mathcal{R}_{\texttt{sol}}$) is a $(k+1)$-ary relation on $\mathcal{P}$ (resp. on $\mathcal{S}$). Then, the inference is based on the following *plausible* inference rule:

$$\frac{\mathcal{R}_{\texttt{pb}}(\texttt{x}^1, \texttt{x}^2, \ldots, \texttt{x}^{k+1})}{\mathcal{R}_{\texttt{sol}}(\texttt{y}^1, \texttt{y}^2, \ldots, \texttt{y}^{k+1})} \quad \text{(with } \texttt{y}^i \text{ a solution of } \texttt{x}^i \text{ for } i \in \{1, 2, \ldots, k+1\}\text{)}$$

The search aims to find k source cases $(\texttt{x}^1, \texttt{y}^1)$, $(\texttt{x}^2, \texttt{y}^2)$, …, $(\texttt{x}^k, \texttt{y}^k)$ such that $\mathcal{R}_{\texttt{pb}}(\texttt{x}^1, \texttt{x}^2, \ldots, \texttt{x}^k, \texttt{x}^{\texttt{tgt}})$ and adaptation consists of solving the constraint $\mathcal{R}_{\texttt{sol}}(\texttt{y}^1, \texttt{y}^2, \ldots, \texttt{y}^k, y)$ with unknown y: a solution of this constraint is a proposed solution $\texttt{y}^{\texttt{tgt}}$ of $\texttt{x}^{\texttt{tgt}}$.

This inference rule can be reinterpreted when $\mathcal{R}_{\texttt{pb}}$ becomes a fuzzy (or gradual) relation, i.e. $\mathcal{R}_{\texttt{pb}}(\texttt{x}^1, \texttt{x}^2, \ldots, \texttt{x}^{k+1})$ may be neither false nor true, but may take intermediate truth values. Then, the inference rule is read as follows:

The more $\mathcal{R}_{\texttt{pb}}(\texttt{x}^1, \texttt{x}^2, \ldots, \texttt{x}^{k+1})$ is, the more plausible $\mathcal{R}_{\texttt{sol}}(\texttt{y}^1, \texttt{y}^2, \ldots, \texttt{y}^{k+1})$ is.

The implemented approaches are based on this "fuzzy interpretation" of the inference rules.

[4] Other variants $p \times q$ could have been considered but the differences between the variants 1×1 and 2×3 in the first experiments have not pushed us towards this direction.

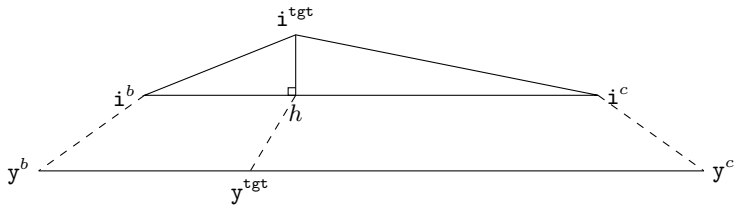


Fig. 2: Some Euclidian geometry for interpolation.

Approximation corresponds to $k = 1$, and $\mathcal{R}_{\mathtt{pb}}$ and $\mathcal{R}_{\mathtt{sol}}$ chosen as "similarity relations". This means that if $\mathtt{x}^s$ is similar to $\mathtt{x}^{\mathtt{tgt}}$ then $\mathtt{y}^s$ is similar to the searched solution $\mathtt{y}^{\mathtt{tgt}}$.

One way to implement this approach is based on the interpretation of the similarity relation as a fuzzy relation, i.e. in selecting the source case $(\mathtt{x}^s, \mathtt{y}^s)$ that is the most similar to $\mathtt{x}^{\mathtt{tgt}}$ and taking $\mathtt{y}^s$ as a solution to $\mathtt{x}^{\mathtt{tgt}}$ (i.e. $\mathtt{y}^{\mathtt{tgt}} = \mathtt{y}^s$). For this purpose, the similarity between the problems is computed thanks to the distance `dist1` between the indexes $\mathtt{i}^s = \mathtt{idx}(\mathtt{x}^s)$ and $\mathtt{i}^{\mathtt{tgt}} = \mathtt{idx}(\mathtt{x}^{\mathtt{tgt}})$: the most the dirtiness index of the images are similar, the most plausible it is that the solution triple of the first image is good for the second one. The approaches implemented are denoted by `approx1x1` and `approx2x3` (approximation approaches with 1×1 and 2×3 dirtiness indexes).

Interpolation corresponds to $k = 2$ and to betweenness relations, following the ideas of interpolative reasoning as presented in [15]. Depending on the choice of the distance functions between dirtiness indexes and between solutions, and on the choice between crisp and fuzzy interpretations of the inference rules, there are several interpolation approaches.

One way to consider it corresponds to the crisp interpretation of the inference rule, the `dist1` distance function between 1×1 dirtiness indexes and the `dist1` distance function between solutions. This means that retrieval selects two source cases $(\mathtt{x}^b, \mathtt{y}^b)$ and $(\mathtt{x}^c, \mathtt{y}^c)$ such that $\mathtt{i}^{\mathtt{tgt}} \in [\mathtt{i}^b, \mathtt{i}^c]$ (if no such pair of cases exist, the interpolation fails). To have the most accurate framing of $\mathtt{idx}(\mathtt{x}^{\mathtt{tgt}})$, the pair of retrieved source cases are chosen according to these criteria:

$$\mathtt{i}^b = \max\{\mathtt{i}^s \mid (\mathtt{x}^s, \mathtt{y}^s) \in \mathtt{CB} \text{ and } \mathtt{i}^s \leq \mathtt{i}^{\mathtt{tgt}}\}$$
$$\mathtt{i}^c = \min\{\mathtt{i}^s \mid (\mathtt{x}^s, \mathtt{y}^s) \in \mathtt{CB} \text{ and } \mathtt{i}^s \geq \mathtt{i}^{\mathtt{tgt}}\}$$

Then, the plausible inference leads to a range $[\mathtt{y}^b, \mathtt{y}^c]_{\mathtt{dist1}}$ for $\mathtt{y}^{\mathtt{tgt}}$, that is,

$$b\ell^{\mathtt{tgt}} \in [b\ell^b, b\ell^c] \qquad w\ell^{\mathtt{tgt}} \in [w\ell^b, w\ell^c] \qquad \gamma^{\mathtt{tgt}} \in [\gamma^b, \gamma^c]$$

where $\mathtt{y}^{\mathtt{tgt}} = (b\ell^{\mathtt{tgt}}, w\ell^{\mathtt{tgt}}, \gamma^{\mathtt{tgt}})$, $\mathtt{y}^b = (b\ell^b, w\ell^b, \gamma^b)$ and $\mathtt{y}^c = (b\ell^c, w\ell^c, \gamma^c)$. This has not been implemented but is considered as future work (see Section 7).

Another way to consider interpolation corresponds to the fuzzy interpretation of the inference rule, the `dist2` distance function between dirtiness indexes, and

the $\texttt{dist2}$ distance function between solutions. The Euclidian distance $\texttt{dist2}$ allows us to use the notion of orthogonal projection of a point on a line. Let $(\texttt{x}^b, \texttt{y}^b)$ and $(\texttt{x}^c, \texttt{y}^c)$ be two source cases and let h be the orthogonal projection of $\texttt{i}^{\texttt{tgt}}$ on the line $(\texttt{i}^b\texttt{i}^c)$ (see Figure 2, for illustration). If h does not belong to the segment $[\texttt{i}^b, \texttt{i}^c]_{\texttt{dist2}}$ or is equal to one of its extremities then this pair of source cases is not a candidate for retrieval. So, in the following, it is assumed that $h \in [\texttt{i}^b, \texttt{i}^c]_{\texttt{dist2}} \setminus \{\texttt{i}^b, \texttt{i}^c\}$. Let $D^{bc} = \texttt{dist2}(\texttt{i}^{\texttt{tgt}}, h)$, the distance between $\texttt{i}^{\texttt{tgt}}$ and the segment $[\texttt{i}^b, \texttt{i}^c]_{\texttt{dist2}}$. $\texttt{i}^{\texttt{tgt}}$ is between $\texttt{i}^b$ and $\texttt{i}^c$ iff $D^{bc} = 0$. The betweenness relation is fuzzified according to the following principle: the lower D^{bc} is, the more $\texttt{i}^{\texttt{tgt}}$ is between $\texttt{i}^b$ and $\texttt{i}^c$. Therefore, the retrieval aims to find a pair of source cases $(\texttt{x}^b, \texttt{y}^b)$ and $(\texttt{x}^c, \texttt{y}^c)$ to minimize D^{bc}. The interpolation principle entails that the proposed solution belongs to $[\texttt{y}^b, \texttt{y}^c]_{\texttt{dist2}}$. However, this interpolation approach proposes a precise value for $\texttt{y}^{\texttt{tgt}}$ by applying the following principle: the closer h is to $\texttt{x}^b$, the closer the solution $\texttt{y}^{\texttt{tgt}}$ proposed for $\texttt{x}^{\texttt{tgt}}$ is to $\texttt{y}^b$. In other words, if h is the barycenter of $(\texttt{i}^b, \alpha)$ and $(\texttt{i}^c, \beta)$ (where $\alpha, \beta \in \mathbb{R}_+^*$) then the proposed $\texttt{y}^{\texttt{tgt}}$ is the barycenter of $(\texttt{y}^b, \alpha)$ and $(\texttt{y}^c, \beta)$, as illustrated in Fig. 2.[5] This has been implemented for the two variants of the dirtiness index, with the implementation names $\texttt{interp1x1}$ and $\texttt{interp2x3}$.

Extrapolation corresponds to $k = 3$ and to analogical proportions on dirtiness indexes and on solutions. Retrieval aims at finding 3 source cases $(\texttt{x}^a, \texttt{y}^a)$, $(\texttt{x}^b, \texttt{y}^b)$ and $(\texttt{x}^c, \texttt{y}^c)$ such that $\texttt{i}^a\texttt{:}\texttt{i}^b\texttt{::}\texttt{i}^c\texttt{:}\texttt{i}^{\texttt{tgt}}$ (the analogical proportion is computed on image dirtiness indexes). Adaptation consists of solving the analogical equation $\texttt{y}^a\texttt{:}\texttt{y}^b\texttt{::}\texttt{y}^c\texttt{:}y$ and giving the solution y as a proposed solution $\texttt{y}^{\texttt{tgt}}$ to $\texttt{x}^{\texttt{tgt}}$.

In order to put this principle into practice, a few remarks can be made.

First, the situations where $\texttt{i}^a\texttt{:}\texttt{i}^b\texttt{::}\texttt{i}^c\texttt{:}\texttt{i}^{\texttt{tgt}}$, i.e. $\texttt{i}^b - \texttt{i}^a = \texttt{i}^{\texttt{tgt}} - \texttt{i}^c$, appear to be rare, so a gradual approach is used by finding the triple of cases for which this relation is best approximated. For this purpose, the measure used consists in computing $\mathcal{AD}(a, b, c, \texttt{tgt}) = \left\|(\texttt{i}^b - \texttt{i}^a) - (\texttt{i}^{\texttt{tgt}} - \texttt{i}^c)\right\|_{\texttt{dist1}}$: the lower $\mathcal{AD}(a, b, c, \texttt{tgt})$ is, the best the approximation is supposed to be (the exact analogy holds iff $\mathcal{AD}(a, b, c, \texttt{tgt}) = 0$). This is inspired by the idea of analogical dissimilarity [9].

[5] The implementation of this approach requires some classical (and tedious) Euclidian geometry recalled in this note. First, the fact that h belongs or not to the segment $[\texttt{i}^b, \texttt{i}^c]_{\texttt{dist2}}$ can be tested by computing the scalar products of $(\texttt{i}^c - \texttt{i}^b)$ by $(\texttt{i}^{\texttt{tgt}} - \texttt{i}^b)$ and $(\texttt{i}^{\texttt{tgt}} - \texttt{i}^c)$: both products are positive iff h belongs to the segment and is different from the extremities of this segment. Second, D^{bc} is computed as follows. The area A of the triangle $\texttt{i}^b\texttt{i}^{\texttt{tgt}}\texttt{i}^c$ can be computed using the Heron formula: $A = \sqrt{p(p-\ell)(p-m)(p-n)}$ where $\ell = \texttt{dist2}(\texttt{i}^b, \texttt{i}^c)$, $m = \texttt{dist2}(\texttt{i}^c, \texttt{i}^{\texttt{tgt}})$, $n = \texttt{dist2}(\texttt{i}^b, \texttt{i}^{\texttt{tgt}})$, and $p = (\ell + m + n)/2$. Then, the length D^{bc} of the segment $[\texttt{i}^{\texttt{tgt}}, h]_{\texttt{dist2}}$ can be computed by $\texttt{dist2}(\texttt{i}^{\texttt{tgt}}, h) = 2A/\ell$. Then, $\texttt{dist2}(\texttt{i}^b, h)$ can be computed thanks to the Pythagorean theorem: $\texttt{dist2}(\texttt{i}^b, h) = \sqrt{\texttt{dist2}(\texttt{i}^b, \texttt{i}^{\texttt{tgt}})^2 - (D^{bc})^2}$. This provides the value of $\texttt{y}^{\texttt{tgt}}$ as a barycenter of $\texttt{y}^b$, $\texttt{y}^c$ with proper weights: $\texttt{y}^{\texttt{tgt}} = \texttt{y}^b + \frac{\texttt{dist2}(\texttt{i}^b, h)}{\texttt{dist2}(\texttt{i}^b, \texttt{i}^c)}(\texttt{y}^c - \texttt{y}^b)$.

Second, with this approach, it may occur that the inferred solution $\mathtt{y}^{\mathtt{tgt}} = (b\ell^{\mathtt{tgt}}, w\ell^{\mathtt{tgt}}, \gamma^{\mathtt{tgt}})$ does not respect the ranges, i.e. $b\ell^{\mathtt{tgt}} \notin [0,1]$, $w\ell^{\mathtt{tgt}} \notin [0,1]$, $b\ell^{\mathtt{tgt}} > w\ell^{\mathtt{tgt}}$, or $\gamma^{\mathtt{tgt}} < 0$. If such situations occur, the extrapolation fails with these retrieved cases and another triple of cases has to be retrieved and adapted.

A naive implementation of this idea would consist in making 3 nested loops ranging on the case base 3 times, which would lead to a complexity in $\mathcal{O}(|\mathtt{CB}|^3)$. A more efficient algorithm has been implemented. It consists first in an offline treatment building a database that stores, for each ordered pair of source cases $((\mathtt{x}^a, \mathtt{y}^a), (\mathtt{x}^b, \mathtt{y}^b))$, the value $\mathtt{i}^b - \mathtt{i}^a$. This offline treatment is in $\mathcal{O}(|\mathtt{CB}|^2)$ but this complexity can be reduced when dealing with larger case bases, by considering, for example, only pairs of similar source cases. Then, given $\mathtt{x}^{\mathtt{tgt}}$, for each $(\mathtt{x}^c, \mathtt{y}^c) \in \mathtt{CB}$, the difference $\mathtt{i}^{\mathtt{tgt}} - \mathtt{i}^c$ is computed and the database is searched for the nearest $\mathtt{i}^b - \mathtt{i}^a$ (the one minimizing $\mathcal{AD}(a, b, c, \mathtt{tgt})$) by a binary search, hence giving the best pair of source cases $((\mathtt{x}^a, \mathtt{y}^a), (\mathtt{x}^b, \mathtt{y}^b))$, given $(\mathtt{x}^c, \mathtt{y}^c)$. So, after this loop on $(\mathtt{x}^c, \mathtt{y}^c) \in \mathtt{CB}$, the best triple of source cases is retrieved and can be adapted. This implementation is in $\mathcal{O}(|\mathtt{CB}| \log |\mathtt{CB}|)$.

The two implementations are named `extrap1x1` and `extrap2x3`.

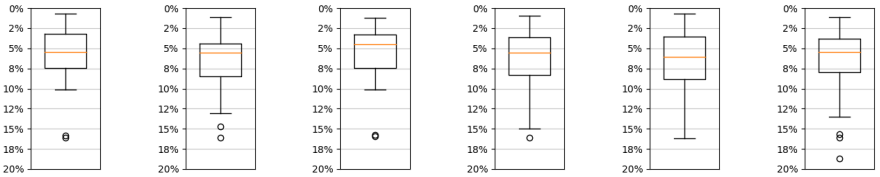
6 Evaluations

The approaches presented in the previous section have been evaluated with a case base and two test sets (Section 6.1). It is composed of an evaluation by experts (Section 6.2), of an automatic evaluation (Section 6.3), and of a qualitative evaluation comparing Georges's output with the output of a selected related work (Section 6.4). The section ends with a discussion on the results of the evaluation (Section 6.5).

6.1 The case base and the test sets

Persée has edited a very large number of documents following the editing process presented in Section 4 and, for each page $\mathtt{x}$ of these documents, the triple $\mathtt{y} = (b\ell, w\ell, \gamma)$ has been stored. The first experiment has shown that with a rather limited case base, the results were good (as the more comprehensive evaluation presented below confirms), so the case base used in the experiments was chosen with a reasonable size of $|\mathtt{CB}| = 11\,166$ and this is the case base used in the three following subsections. Another set of 19 cases has been selected for evaluation by a human expert (see Section 6.2).

Now, as explained before, the triple $\mathtt{y}$ associated to an image $\mathtt{x}$ has been chosen by experts on an image of the same document as $\mathtt{x}$, not necessarily on $\mathtt{x}$ itself: the result is considered good (since it is validated by further editing steps) but may be less good than the one the expert would have chosen on $\mathtt{x}$, and the information about which page of each document was used by the expert has not been retained. For the automatic evaluation (see Section 6.3) a set of 39 cases have been defined by the expert as gold standards: they constitute pairs $(\mathtt{x}, \mathtt{y})$ where $\mathtt{y}$ was determined on the image $\mathtt{x}$ itself. Such cases are called "ideal cases" in the following: they are based directly on the expert's assessment.



`approx1x1` (13 ms) `approx2x3` (10 ms) `interp1x1` (13 ms) `interp2x3` (40 ms) `extrap1x1` (35 ms) `extrap2x3` (56 ms)

Fig. 3: Results of the automatic evaluation, with computing times per target problem (computed on a simple laptop).

6.2 An evaluation by a human expert

Let `TS` be the set of 19 cases selected for this evaluation (`TS` is disjoint from `CB`). An interface has been implemented for this evaluation that consists, for each $(\mathtt{x},\mathtt{y}) \in \mathtt{TS}$, in presenting to the expert the original image $\mathtt{x}$ and 7 cleaned images presented in a random order: the image cleaned by the edited process of Persée (i.e. $\mathtt{clean}(\mathtt{y},\mathtt{x})$, thus a good solution, but in general not an ideal solution) and the 6 images $\mathtt{clean}(\mathtt{y}^g,\mathtt{x})$ for the 6 approaches implemented in Georges. The expert had the possibility to compare images by pairs, with a possibility of zoom to see the details, and associate to each 7 cleaned images a mark in the scale $\{1 = \text{very bad}, 2 = \text{bad}, 3 = \text{good}, 4 = \text{very good}\}$.

The results are as follows, where $a \pm \sigma, b$ means that the average value is a with a standard deviation of σ and b images of this approach were the best ones in the series:

Persée	`approx1x1`	`approx2x3`	`interp1x1`	`interp2x3`	`extrap1x1`	`extrap2x3`
$2.63 \pm 1.07, 3$	$2.88 \pm 0.99, 3$	$2.84 \pm 1.01, 3$	$2.95 \pm 0.85, 2$	$3.21 \pm 0.98, 5$	$2.58 \pm 0.96, 1$	$2.42 \pm 1.12, 2$

6.3 An automatic evaluation

The automatic evaluation aims at comparing the different implemented approaches, using the 39 ideal cases as test sets. For each ideal case $(\mathtt{x},\mathtt{y})$, the target problem is set to $\mathtt{x}$ ($\mathtt{x}^{\mathtt{tgt}} = \mathtt{x}$) and for each approach, the predicted solution $\mathtt{y}^{\mathtt{tgt}}$ is compared to the solution $\mathtt{y}$. Now, this comparison cannot be done only on triples, since two different triples may clean in the same way a given image. Therefore, this comparison is done by comparing directly the resulting images, i.e. by computing $\mathtt{dist1}(\mathtt{clean}(\mathtt{y}^{\mathtt{tgt}},\mathtt{x}^{\mathtt{tgt}}),\mathtt{clean}(\mathtt{y},\mathtt{x}^{\mathtt{tgt}}))$ and normalize it on a $[0\%, 100\%]$ scale (0% stands for identical images).

Figure 3 presents the result of this evaluation using box plots, as well as indicative computing times per target problem. It seems that `interp1x1` is the best one, but the differences from the other approaches is not very significant.

(a) Original image. (b) GEORGES's cleaning. (c) [7]'s cleaning.

Fig. 4: Comparing cleanings of GEORGES and [7] on an example.

6.4 A qualitative comparison with the work of [7]

As presented in Section 3, a system close to GEORGES, in terms of expected output, is the one described in [7]. One advantage of this previous work, compared to GEORGES, is that it is language independent and could be used successfully for documents written for any language, whereas GEORGES requires a case base of text for the same language and an OCR analysis for situating a frequent word (or letter) of this language in text images.

Now, we suspected that the approach presented in this previous work would be harmful for non-textual elements of a page, such as illustrations. To test this hypothesis, this algorithm has been (re-)implemented and compared to GEORGES on several documents containing texts and illustrations and this hypothesis was confirmed. Figure 4 illustrates this (using the approach `interp1x1`).

Other comparisons of GEORGES with cleaning and binarization tools have been carried out and have led to the same conclusion: GEORGES outperforms these generic tools especially on pages combining texts and illustrations. Indeed, GEORGES has benefited from the specific context of images of French texts with potential illustrations. The idea of separating the treatments of texts and illustrations has been rejected because of the inexact segmentation of illustrations and because of the discontinuity of treatment within the same page.

6.5 Discussion

Human evaluation has shown that GEORGES behaves nicely and proposes solutions that are in the average better than what is proposed in the case base $\texttt{CB}$. This is quite surprising, but it can be explained by the fact that although the cases of $\texttt{CB}$ are good (in the sense that they are the values accepted by experts after the execution of the entire editing process), they are not ideal cases in general. For example, if the approximation approach is used, given a case $(\texttt{x}, \texttt{y})$ used in the test set, the closest case $(\texttt{x}^s, \texttt{y}^s) \in \texttt{CB}$ may be taken from an image of a different document. Therefore, the benefits of the automatic CBR approach for cleaning text images are (1) a gain in expert working time, as expected (a few hours per month) and (2) an improvement of the quality of the cleaning process, which has been a surprise.

Automatic evaluation needed a comparison with a gold standard test set, with only ideal cases. Its aim was to compare the different approaches of GEORGES. Now, the result has not shown a clear winner. This is consistent with the human

evaluation, where the best approach was not always the same one. This suggests that a combination of these approaches could be profitable and this is practically feasible, since the approaches take reasonable computing time.

7 Conclusion

This article has presented an application of CBR to the cleaning of images of French texts, using the approximation, interpolation, and extrapolation approaches to CBR and an indexing of images by a dirtiness index. From an application point of view, this work has two main benefits. First, it has transformed a semi-automatic task into an automatic task saving this way several hours per month. Second, it has improved the quality of the result. Indeed, the semi-automatic approach was applied for one page by document (using the same parameters for cleaning its other pages), whereas the automatic approach has made it possible to treat every individual pages, which has resulted in better solutions as the evaluation has shown.

Beyond this application, this work can be used for other applications where a semi-automatic work is based on the choice of parameters by a human expert: this choice could be suggested by a CBR system similar to Georges to lighten the workload of the expert. The remaining difficulty to apply Georges's principle is to build a relevant index on the problems.

A first direction of future work is on the definition of a new dirtiness index: currently, it is based on averages of grayscale values of the occurrences of "de" on the image. It would be interesting to examine how taking into account the standard deviation of "de" occurrences in the image can be used.

Another direction for future work for the development of Georges (and similar CBR applications) is the design of a combination of the different proposed approaches. An obvious approach to do so is to take the average of the solutions proposed by the different approaches.

A less obvious way to do so is to use non-fuzzy approaches for approximation and interpolation. For the latter, the idea has already been presented in Section 5.3: it leads to plausibly inferring that the solution $\mathtt{y}^{\mathtt{tgt}}$ belongs to a range $[\mathtt{y}^b, \mathtt{y}^c]_{\mathtt{dist1}}$. A similar idea for approximation can be to considered: if $\mathtt{x}^s$ and $\mathtt{x}^{\mathtt{tgt}}$ are similar then this gives for the solution $\mathtt{y}^{\mathtt{tgt}}$ a solution range by the set of $\mathtt{y} \in \mathcal{S}$ whose similarity to $\mathtt{y}^s$ is constrained by the distance between $\mathtt{x}^s$ and $\mathtt{x}^{\mathtt{tgt}}$: this idea has been presented in [3] and further developed in [6]. In both approaches, a range is plausibly inferred for $\mathtt{y}^{\mathtt{tgt}}$ and this range could be used in order to improve the combination of the current approaches. An first idea to be explored is to use these ranges as constraints, by keeping only the solutions that are consistent with these ranges.

Finally, the automatic experiment of Section 6.3 has been rerun with a case base of 1 000 cases, chosen randomly in the case base of 11 166 cases, without significant decrease in result. This suggests that a case base maintenance (see e.g. [16]) lowering the number of cases while maintaining the quality of the results should be beneficial.

References

1. Aisbett, J., Gibbon, G.: A general formulation of conceptual spaces as a meso level representation. Artificial Intelligence **133**(1-2) (2001) 189–232
2. Chaki, N., Shaikh, S.H., Saeed, K. In: A Comprehensive Survey on Image Binarization Techniques. Springer (2014)
3. Dubois, D., Hüllermeier, E., Prade, H.: Flexible control of case-based prediction in the framework of possibility theory. In Blanzieri, E., Portinale, L., eds.: Advances in Case-Based Reasoning — Proc. of the 5th Eur. workshop on case-based reasoning (EWCBR'2k). LNCS 1898. Springer (2000) 61–73
4. Ficet-Cauchard, V., Porquet, C., Revenu, M.: CBR for the reuse of image processing knowledge: a recursive retrieval/adaptation strategy. In: Case-Based Reasoning Research and Development: Third International Conference on Case-Based Reasoning, ICCBR-99 Seeon Monastery, Germany. (1999)
5. Gärdenfors, P.: Conceptual spaces: The geometry of thought: Cambridge (2000)
6. Hüllermeier, E.: Credible Case-Based Inference Using Similarity Profiles. IEEE Transaction on Knowledge and Data Engineering **19**(6) (2007) 847–858
7. Kavallieratou, E., Antonopoulou, H.: Cleaning and enhancing historical document images. In: Advanced Concepts for Intelligent Vision Systems: 7th International Conference, ACIVS 2005, Antwerp, Belgium. (2005)
8. Lieber, J., Nauer, E., Prade, H., Richard, G.: Making the best of cases by approximation, interpolation and extrapolation. In: ICCBR 2018 - 26th Int. Conf. on Case-Based Reasoning. (2018)
9. Miclet, L., Bayoudh, S., Delhay, A.: Analogical dissimilarity: definition, algorithms and two experiments in machine learning. Journal of Artificial Intelligence Research **32** (2008) 793–824
10. Nandy, M., Saha, S.: An analytical study of different document image binarization methods. arXiv preprint arXiv:1501.07862 (2015)
11. Perner, P.: Why case-based reasoning is attractive for image interpretation. In: Case-Based Reasoning Research and Development: 4th International Conference on Case-Based Reasoning, ICCBR-2001 Vancouver, BC, Canada, July 30–August 2, 2001 Proceedings, Springer (2001) 27–43
12. Perner, P., Holt, A., Richter, M.: Image processing in case-based reasoning. The Knowledge Engineering Review **20**(3) (2005) 311–314
13. Prade, H., Richard, G.: From analogical proportion to logical proportions. Logica Universalis **7**(4) (2013) 441–505
14. Riesbeck, C.K., Schank, R.C.: Inside Case-Based Reasoning. Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey (1989) Available on line.
15. Schockaert, S., Prade, H.: Interpolative and extrapolative reasoning in propositional theories using qualitative knowledge about conceptual spaces. Artificial Intelligence **202** (2013) 86–131
16. Smyth, B., Keane, M.T.: Remembering to forget. In: Proc. of the 14th Int. Joint Conf. on Artificial Intelligence (IJCAI'95), Montréal (1995)
17. Tensmeyer, C., Martinez, T.: Historical document image binarization: A review. SN Computer Science **1**(3) (2020) 173–198