Detection of Figure and Caption Pairs based on Disorder Measurements

Claudie Faure [a], Nicole Vincent[b]
[a] CNRS-LTCI, TELECOM-ParisTech, 46 rue Barrault, 75634 Paris, Cedex 13, France
[b]LIPADE - Université Paris Descartes, 45, rue des Saints-Pères, 75270 Paris Cedex 06, France

ABSTRACT

Figures inserted in documents mediate a kind of information for which the visual modality is more appropriate than the text. A complete understanding of a figure often necessitates the reading of its caption or to establish a relationship with the main text using a numbered figure identifier which is replicated in the caption and in the main text. A figure and its caption are closely related; they constitute single multimodal components (FC-pair) that Document Image Analysis cannot extract with text and graphics segmentation. We propose a method to go further than the graphics and text segmentation in order to extract FC-pairs without performing a full labelling of the page components. Horizontal and vertical text lines are detected in the pages. The graphics are associated with selected text lines to initiate the detector of FC-pairs. Spatial and visual disorders are introduced to define a layout model in terms of properties. It enables to cope with most of the numerous spatial arrangements of graphics and text lines. The detector of FC-pairs performs operations in order to eliminate the layout disorder and assigns a quality value to each FC-pair. The processed documents were collected in medic@, the digital historical collection of the BIUM (Bibliothèque InterUniversitaire Médicale). A first set of 98 pages constitutes the design set. Then 298 pages were collected to evaluate the system. The performances are the result of a full process, from the binarisation of the digital images to the detection of FC-pairs.

Keywords: Historical documents, Figure-Caption pairs, Spatial reasoning, Spatial Disorder, Visual Disorder

1. INTRODUCTION

Information retrieval methods are specialised to explore the textual content of documents or the visual content of image databases. When figures are inserted in the main text of documents, such as technical books or journals, then both visual and textual modalities are involved to mediate the document content. The meaning of a figure cannot be fully understood without the reading of the related text parts. When a caption contains a figure identifier, it can be used to find text parts related to the figure by searching occurrences of the identifier in the main text. A caption may also contain a text explaining what to see in a figure and it is difficult to fully understand the caption in the absence of its visual reference. As shown in [12, 3, 11] for photographs associated with captions, the integration of visual and linguistic information leads to better results in classification and understanding of the photographs than the analysis of a single modality. Textual and visual modalities cannot be separated in figure and caption pairs (FC-pairs) which constitute multimodal components.

Our purpose is to detect FC-pairs in historical documents printed in the 19th Century in order to assist the archivists in the indexing and storage of digitised books. Till now, many tasks are performed manually. In medic@, the digital library of the BIUM (Bibliothèque InterUniversitaire Médicale [13]), the archivists select and store the pages containing FC-pairs in an image database and indicate in the book summaries which pages contain FC-pairs along with the text of each caption.

The detection of FC-pairs implies a prior segmentation into graphics and text components. Captions may be detected during the logical structure analysis when the physical components are labelled, see [13] where the logical labels of 17 surveyed systems are given. Few systems perform a goal-driven process to detect figures and captions enabling to focus the attention of the logical analysis on the neighbourhood of visual components [7, 8]. A goal-driven process has the advantage to ignore irrelevant parts of the documents during the analysis and to avoid a complete logical analysis of the document. In [8], the FC-pairs are extracted from digital documents encoded in pdf format exhibiting text and images components. The main problem solved in [8] is to discriminate the caption blocks from the other text blocks of the

document and to find the right image and caption associations. It is said, but not precisely described, that standard publishing rules for the layout and typographic attributes are involved to search the associations with a constraint satisfaction method. The goal of this work is to store the visual parts into an image database and to exploit the captions as annotations to improve information search. In [7] captions are detected in digitized patents. In these documents, the figures are gathered in the pages of the figure section and not displayed in the pages containing the main text. The pages of the figure section are a mixture of text (labels and captions) and graphics. The main problem is to discriminate text strings and graphics for which the authors propose a method relevant for graphic documents. Then, the text strings are sent to an OCR to discriminate labels and captions (the figure identifiers). The association of graphics and captions is not performed, the FC-pairs are not segmented. The relationship between text and figures is mediated by the figure identifiers. The goal is to provide a user-friendly navigation in the digital version of the patent. A figure identifier present in the main text and in the figure section are linked, enabling a direct access between the page where a figure is displayed and the page where the main text describes the figure.

The detection of FC-pairs in historical documents imply to solve problems that are not encountered in [7, 8]. Unlike in [8] the segmentation into text and graphic components is needed. Unlike in [7], the figures are built from the detected graphics components. After the text/graphics segmentation, graphic components are not always the expected figures (a figure may be composed of several disjoint graphics for example). The composition and printing techniques as well as the lack of precise publishing rules in the 19th Century lead to a broad range of layouts and prevent the definition of safe rules to perform an early detection of the figures from the graphic components. Moreover, a figure bounding box may also contain its caption. Our solution is to postpone the final figure segmentation during the FC-pairs detection process. Another problem encountered in the 19th Century books is the unpredictable direction of the text lines, a page may contain vertical and horizontal text lines. This led us to propose a method to simultaneously detect vertical and horizontal text lines [2]. These text lines were automatically sorted in order to associate a set of caption line candidates with each graphic component. We cannot assume, as in [8], that captions are horizontal text lines and located above or under the figures, they may be vertical text lines located on the right or on the left of the figure. This increases the number of candidates for figure and caption associations. It has been shown that word spotting reduces greatly the number of caption line candidates when a figure identifier (such as "Fig.") is found and it may also detect true caption lines that were not selected in the candidate set [5].

In the present study, the detection of the FC-pairs is performed using spatial reasoning involving text lines and graphic components. A layout model is needed to drive the exploration of page components towards a meaningful result. As noticed in [11], much of the work in logical structure analysis assumes that physical layout analysis has already been performed and do not handle inaccurate inputs. The knowledge used to label the physical components is represented by descriptive models based on geometrical and spatial relationships between components. Descriptive models based on publishing rules are able to drive efficiently the detection of FC-pairs [8] or to propose realistic reading orders [1] when the components are pre-segmented but they do not tolerate imperfect component segmentation. Instead of a descriptive model, we propose a layout model based on properties related to the order and regularity requirements of layout design. If the spatial arrangement of the components in a page appears unorganised, reading becomes a fastidious exploration of the page to reconstruct meaningful relationships. A clean and organised layout is not only an aesthetic requirement [4] but also the way to ensure an easy and efficient reading. The detection of FC-pairs is aimed at finding a disorder-free spatial arrangement of graphics and related text components in a page. The disorder is represented by a set of properties defining the layout model. This model fits most of the component arrangements encountered in the data and it involves a reduced context; segmentation and labelling of all the page components is not necessary.

The full system achieves successive stages of processing. The text and graphic components are first discriminated. Then text lines are extracted and a set of caption line candidates is selected for each graphic component. The FC-pairs' detection process perform a set of operations to eliminate the layout disorder and to reach a final decision.

The next section gives a brief description of the segmentation into basic components. Section 3 gives the rules used to select a set of caption line candidates and the definition of the uncertainty value characterising a pair of associated graphics and text line. The disorder measurements and the layout model are introduced in Section 4. Section 5 describes the steps of the process aimed at eliminating the layout disorder. The results are given in Section 6.

2. BASIC COMPONENTS

The images are first binarised [6] to obtain the connected components (CCs). Size and shape criteria are used to sort the CCs. The largest CCs are interpreted as Graphics and labelled CCG. The shape is considered to detect graphic lines and

frames. The CCs in contact with the image borders are Noise. The remaining components are processed to detect the text lines. The segmentation into basic components is fully described in [2]. In the following, CCG will denote either the CCG or its bounding box according to the context.

The image analysis was performed to prepare the detection of FC-pairs in books printed in the 19th Century. We recall here how the anticipation of the detection of FC-pairs influenced image analysis. The pages may contain horizontal and vertical text lines and it is not possible to predict the direction of the caption lines from the dominant direction in a page. This led us to propose a method to simultaneously detect horizontal and vertical text lines. We also observed that captions or even FC-pairs might be located inside CCG bounding boxes. For avoiding the loss of this information, text line detection was performed both inside and outside the CCGs. Graphics are often not fully connected, merging overlapping CCGs is a way to improve the segmentation of graphics. As a drawback, a figure located inside the bounding box of another figure disappears when merging the CCGs. Therefore, the initial CCGs are not discarded, they are included in the set of graphic components with the CCGs resulting from merge operations.

Segmentation into basic components is not error-free. Errors may be recovered during the detection of the FC-pairs, this is the case for the merged text lines in Fig. 2e,f. The most damaging error is when the confusion of graphics with text prevents the detection of small figures, in this case the FC-pairs cannot be found (see Fig. 2g).

3. FIGURE AND CAPTION PAIRS

Typography and layout are chosen to help the reader to rapidly access the information and to avoid ambiguities. Proximity, alignment and similarity are the properties enabling the human reader to perform perceptual grouping involved in the detection of text components at several levels (eg. words, lines, paragraphs, columns). The text line segmentation method described in [2] was based on perceptual grouping. In the present case, the reader has to associate components (figure and caption). Association cannot be considered as perceptual grouping as defined by the gestalt laws. The components to be associated differ by their modalities (text and graphics); therefore the similarity condition is not satisfied. Sequences of regularly spaced symbols or lines determine the visual saliency of text lines and paragraphs. There is no such regular spatial arrangement of components in FC-pairs. Nevertheless, readers perceive properties enabling them to detect the implicit link existing between a figure and its caption. Association appears to be a high level perceptual process involving a priori domain knowledge and decisions.

The association of components is strongly determined by proximity. As noticed by [8], an implicit link between neighbouring components is detected by the readers when their relative position may be expressed such as: “close to”," "on top of", “to the right of”, "under" ... A figure and its caption are always neighbours but lines of the main text may be closer to the figure than the caption lines. Other spatial relationships help to disambiguate the association. In all the cases, the figure and caption projections overlap on the x or y axis and the link between them is more obvious if their centres are aligned (along a vertical or a horizontal direction).

Another property of the layout is the compactness. The rectangular bounding box of a FC-pair is filled with the two components, a figure and a caption, so as to avoid white zones in the layout, which are wasted space and could be interpreted as a missing component by the reader. Compactness is satisfied when the main direction of the caption lines is parallel to the closest border of the figure: horizontal text lines are above or under the figure while vertical text lines are on the left or on the right of the figure.

A set of caption line candidates is searched for to initiate the FC-pair's detection process. The text lines are sorted according to the proximity, overlapping projections and compactness. For each CCG border, the closest text line is selected. A selected text line becomes a caption line candidate if it is the closest line to the CCG or if the centres of the line and of the CCG are vertically or horizontally aligned. The CCGs and their associated caption candidates constitute a set of Graphic and Text pairs (GT-pairs) which are stored in a data structure. At this step (Step 0 in Fig. 1), the GT-pairs may be the true FC-pairs in some pages (see the Results section). For most pages, the process is carried on to reach the final decision. Fig. 1 shows the steps of the detection of FC-pairs, with the conditions required to activate the related actions. The process starts (Step 1 in Fig. 1) with a decision based on the quality of the GT-pairs: let G be a graphic component, the GT_k-pair with the best quality is selected among $\{GT_1, GT_2, ... GT_n\}$, $n > 1$.

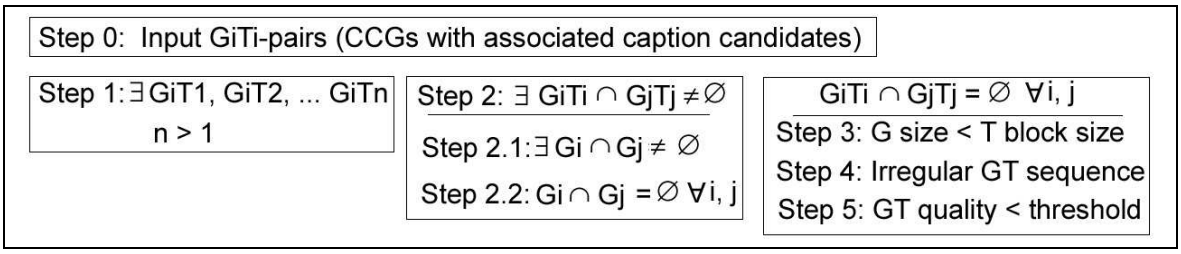


Figure 1. The steps of the FC-pairs detection process with the conditions required to activate the actions.

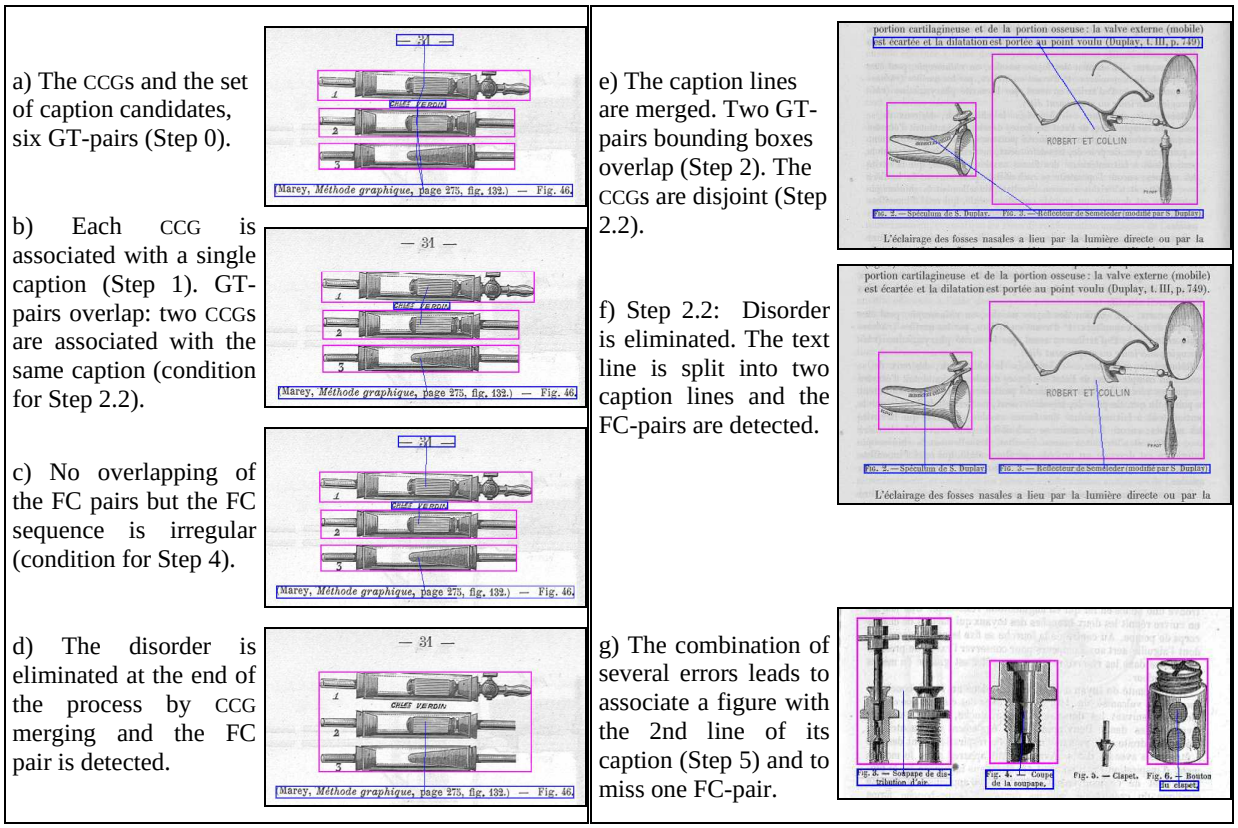


Figure 2. Examples of GT-pairs at different steps of the process (steps are given in Fig. 1)

To evaluate the quality of a GT-pair, a set of five features V_i is introduced. The details of the definitions of V_1, V_2 and V_3 are given for a horizontal text line above or under the CCG, they are easy to adapt for a vertical text line on the left or the right of the CCG.

- V_1 encompasses the proximity property. V_1 increases with the distance between the text line and the closest border of the CCG bounding box (*dfc*). When *dfc* is greater than twice the height of the figure (*hf*), V_1 is equal to 100, its maximal value.

$$V_1 = Min(100, (100 * dfc)/(2 * hf))$$

- V_2 is function of the distance between the centre of the text line and the centre of the CCG (*dcc*). When *dcc* is greater than the width of the figure (*wf*), V_2 is equal to 100, its maximal value.

$$V_2 = Min(100, (100 * dcc)/(wf))$$

- V_3 is related to the compactness, it is the length of the text line parts which overtake the figure bounding box on the left side (*dl*) and on the right side (*dr*). When (*dl* + *dr*) is greater than twice the width of the figure (*wf*), V_3 is equal to 100, its maximal value.

$$V_3 = Min(100, (100 * (dl + dr)) / (2 * wf))$$

- V_4 is the confidence value calculated for each text line during image analysis. V_4 is a linear combination of the number of CCs in the text line, the ratio of the average inter-CC distance and the maximal distance, the ratio of the average height of the CCs and the maximal height, the ratio of the sum of the CC surfaces inside the line and the surface on the line bounding box.
- V_5 is a binary value chosen to express the preference for relative positions of a figure and its caption. The positions "under" and "on the right" are more often encountered than "above" and "on the left".

The final uncertainty value of a GT-pair is given by the weighted sum of the V_is:

$$UGT = (0.25 * V_1) + (0.4 * V_2) + (0.1 * V_3) + (0.1 * V_4) + (0.15 * V_5)$$

The UGT values are calculated for the GT-pairs defined by a CCG and each of its caption candidates. The lowest UGT values characterise the GT-pairs of best quality. The first level of decision associates a single caption candidate with each CCG. The data structure keeps in memory the whole set of GT-pairs and the history of the process. Non-selected GT-pairs are inhibited, never eliminated. Fig. 2a,b shows the GT-pairs before and after the first decision

4. DISORDER MEASUREMENTS

The GT-pairs obtained after the first decision constitute a set of many different spatial arrangements of CCG and caption candidates. For each arrangement, it was easy to understand which operations to perform to extract the true FC-pair. The question was: how to identify an arrangement in order to perform the relevant operations? The requirement for organised layout led us to represent the different arrangements in terms of properties based on order, regularity and consistency. The set of properties constitutes the layout model. FC-pairs' detection becomes a disorder elimination process driven by the layout model. The question arising now is the definition of the ordering properties and therefore, of the disorder.

The aesthetic criteria introduced by [4] are based on a set of measurements such as the misalignment of the borders (or the centres) of the components, irregular spacing between components or spurious white zones. They may be interpreted as disorder measurements. The V_2 and V_3 features, defined in the previous section, encompass the aesthetic alignment criteria related to the border and centre alignments. Following [4], we extend the measurement of the spatial disorder to several components and following [1] the logical labels of the components are considered in the layout model. Spatial disorder is represented by two Boolean variables that are defined according to spatial relationships:

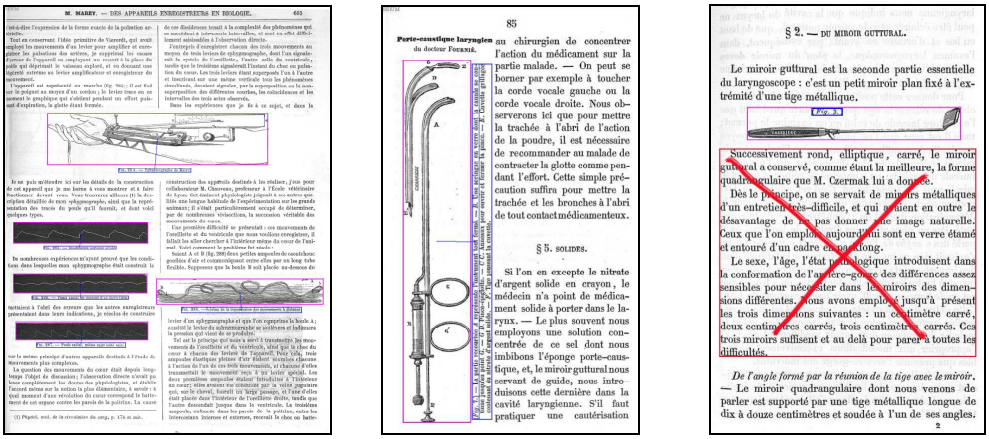
- SD1 is true if components have overlapping bounding boxes. The components that will be considered are the CCGs and the GT-pairs.
- SD2 is true if the sequencing of figure and caption is irregular in a page; meaning that the relative positions of figure and caption are not the same for all the GT-pairs while the caption line directions are identical (horizontal or vertical). Examples are given in Fig. 2c, e.

Visual disorder variables are introduced to only accept CCGs and GT-pairs for which the most represented modality in their bounding box is graphics, and not text.

- VD1 is true for a CCG if text lines occupy a surface of the CCG greater than the surface occupied by graphics.
- VD2 is true when the height of the caption block is greater than the height of the figure for horizontal caption lines (width replaces height for vertical caption lines). We consider that the caption adds information to the figure but the graphics modality dominates in FC-pairs.

5. DETECTION OF FC-PAIRS

The layout model defined in the previous sections is efficient to cope with most cases and to enable a model driven processing. Eliminating the disorder implies to perform a set of actions when conflicts with the layout model are found, i.e. when a disorder variable is true. The conflicts are prioritised to reduce the amount of processing and to take advantage of a modular system in the design and the evaluation stages. At each step of the process (Fig. 1), the actions are performed on the basic components and/or on the data structure to update the set of GT-pairs. Step 0 is the selection of caption line candidates, Step 1 associates a single caption with each CCG, both steps are described in Section 3.



a b c

Figure 3. Examples of detected FC-pairs.

Step 2 is activated when SD1 is true for two GT-pairs: G_1T_1 and G_2T_2 have overlapping bounding boxes. The overlapping of CCGs is first considered in step 2.1. The possible action to reduce the disorder is to merge the CCGs into CCGm. The uncertainty values of the GT-pairs (CCGm- T_1) and (CCGm-T_2) are calculated to select the best GT-pair. The data structure is updated: the GT-pair involving the non-selected caption is inhibited, CCGm replaces CCG in the other GT-pair, and the uncertainty value is updated.

Step 2.2. processes overlapping GT-pairs for which the CCGs are disjoint. Fig. 2 shows examples: two CCGs have a caption in common (Fig. 2b), the text line associated with each CCG overlaps the other CCG on the x direction (Fig. 2e). - Step 2.2. starts with an attempt to obtain two disjoint GT-pairs. Two actions are possible: finding a new caption to associate with one of the CCG (Fig. 2c) or splitting a caption line (Fig. 2f). Splitting a text line, which was merged erroneously during image analysis, is conditioned: the location of the cut is between the two CCGs and corresponds to the maximal length interval between two successive CCs in the text line (see an example of split line in Fig. 2e,f). If disjoint GT-pairs cannot be built, then the CCGs are merged if the resulting CCGm satisfies the condition of visual disorder (VD1 is false). This condition prevents erroneous CCGm for which parts of the main text would be included in its bounding box. Once an action is performed, the GT-pairs are updated. It may happen that merging leads to include the caption lines into the CCGm, in this case, a search for a new caption line outside CCGm is performed.

The next steps are activated when SD1 is false (the G_iT_is are not overlapping). In Step 3, the caption blocks are built from the caption lines for each GT-pair and the value of VD2 is calculated. If VD2 is true for a GT-pair, then a search for a new caption line is performed. If a line satisfying the block size condition (VD2) is found then the GT-pair is updated. This condition avoids confusing a line of the main text with a caption line. Fig.3c shows an example for which the figure was associated with a line of the main text (under the figure), the block size condition inhibits this GT-pair (the caption block is crossed in Fig. 3c).

Step 4 is activated when SD2 is true. Vertical sequences and then horizontal sequences are sought. A vertical sequence of GT-pairs is found if the CCGs are disjoint along the vertical direction. An irregular sequence is shown in Fig. 2c with the relative positions: Above (G_1, T_1) – Above (G_2, T_2) – Under (G_3, T_3). A search for a new caption line is performed to obtain a regular sequence. If the search fails, then the CCGs are merged into CCGm and the best GT-pair (CCGm -Ti) is selected.

Step 5 is the final step where the decision for the FC-pairs is made according to the UGT values. GT-pairs with low UGT values become FC-Pairs. For the others, a better solution is sought. If a GT-pair has a high UGT value, then the bounding box of its CCG is expanded by including the caption line. A new caption line is looked for in the neighbourhood of the expanded CCG. The new GT-pair must have a UGT lower than the current one to be accepted. When a better caption line is not found outside an expanded CCG, a caption line is sought inside the CCG. The condition to activate the search for a better caption line implies to define thresholds for the UGT of the GT-pairs, a threshold t1 for the preferred positions (Under(G, T) and On the Right(G, T)) and t2 for the others (On the Left(G, T) and Above(G, T)) with t1 > t2. Fig. 3d gives an example where the caption is found inside the CCG.

During the detection of FC-pairs, the actions performed are conditioned in order to avoid the introduction of new sources of disorder and cyclic decisions. For example, the search for a better solution in Step 5 cannot destroy the regular sequences of figures and captions already established in step 4.

6. RESULTS

The method was developed using a set of pages collected in medic@. The rule we adopted to build this set was to select the first ten FC-pairs in ten books. Because some books have less than ten figures we obtain a design set of 117 FC-pairs gathered from 12 books. Then, a new set of 298 pages (470 FC-pairs) was processed for the evaluation of the system. The results are given for the design set and the evaluation set.

The number of steps which are needed to reach the final decision for a page depends on the processed data. Table 1 shows the number of pages for which the final decision is obtained when the caption candidates are selected (Step 0) or after Step 2 when the GT-pairs are disjoint in a page or when the further steps of processing are activated. The number of correctly detected FC-pairs (OK) and the number of Accepted FC-pairs are given for these decisions steps. A FC-pair is Accepted when the figure is associated with the figure title or with a descriptive caption, and not with the caption line containing the figure identifier.

Table 1. Number of pages and number of FC-pairs in the data sets. Number of final decisions and number of correct decisions at three steps of the process.

	# Pages	# FC pairs	Final decision steps : # Pages / # OK FC-pairs / # Accepted FC-pairs		
			Caption candidates	End of step 2	End of process
Design set	98	117	9 / 10 / 0	73 / 76 / 5	16 / 13 / 1
Evaluation set	298	470	28 / 28 / 2	226 / 294 / 3	44 / 63 / 3

The results are given in Table 2. Three kinds of errors are reported:

- Errors of type 1 occur when the spatial arrangement of the FC-pairs matches the layout model but the system fails to detect it. Errors of type 1 may be the consequence of previous errors in the Graphics/Text segmentation performed by the system. Another frequent error of type 1 is the association of a figure with a line of the main text for which the UGT of the GT-pair is lower than the UGT of the true FC-pair (the difference of the UGT values is usually low).
- The errors of type 2 correspond to spatial arrangements which do not match the property-based layout model. An error of type 2 occurs when a true FC-pair is included in the bounding box of another FC-pair (SD1 is true) or when the true FC-pairs in a page do not constitute ordered sequences (SD2 is true).
- FC-pairs are added. This happens when graphic lines in skewed documents or large title characters are confused with CCGs. Text lines are associated with these CCGs to produce false FC-pairs.

Table 2. Results obtained for the detection of the FC-pairs

#FC pairs	Ground truth	OK	Accepted	Errors type 1	Errors type 2	Added
Design set	117	99 (85%)	6 (5%)	9 (7%)	3 (3%)	0
Evaluation set	470	385 (82%)	8 (2%)	55 (12%)	22 (5%)	6

Table 3 gives the values of Precision, Recall and F-score calculated with to the following formulas:

Precision 1 = Number of "OK" FC-pairs / Number of detected FC-pairs
Precision 2 = Number of "OK" FC-pairs + number of "Accepted" FC-pairs / Number of detected FC-pairs
Recall 1 = Number of "OK" FC-pairs / Number of FC-pairs in the ground truth
Recall 2 = Number of "OK" FC-pairs+number of "Accepted" FC-pairs / Number of FC-pairs in the ground truth
F-score 1 = 2 P1.R1 / (P1 + R1) F-score 2 = 2 P2.R2 / (P2 + R2)

Table 3. Precision, Recall and F-score

	Precision 1	Precision 2	Recall 1	Recall 2	F-score 1	F–score 2
Design set	0,88	0,93	0,85	0,9	0,86	0,91
Evaluation set	0,85	0,87	0,82	0,84	0,84	0,85

7. CONCLUSION

The results show the current performances of a system aimed at extracting the FC pairs from digitised documents printed in the 19th Century. The processed documents contain many figure styles and a broad range of spatial arrangements of FC-pairs in a page. A small set of disorder measurements was introduced to define a layout model in terms of properties. This model fits most of the encountered cases and does not need a full description of the spatial arrangement of the page components to drive the detection of the FC-pairs.

We are focussing information retrieval in historical documents. Nevertheless, our system has been run on several pages of contemporary scientific papers containing FC-pairs in order to give evidence of what is specific to historical document in our system. The colour is not handle in our system, the figures inserted in the main text pages of 19th Century books are not in colour. Pages containing vertical and horizontal text lines are too frequent in 19th Century books to assume that text lines are horizontal, an acceptable assumption for contemporary documents. The main limitation of our system when processing contemporary documents is the detection of graphic components. The 19th Century documents are not very sensitive to the weakness of the graphics detection method but it prevents to successfully detect FC-pairs in contemporary scientific documents when a figure is a mixture of small geometrical objects, graphic lines and text lines, as it is often the case. The current conventions in scientific documents are to place the figure captions under the figures and the table captions above the tables. In our system, an efficient step (step 4) of the process builds regular sequences of graphics and associated text lines (for example, horizontal caption lines are all under the figures in a page). This step is a source of errors when at least a table and a figure are in the same page of a contemporary scientific paper.

Results will be improved with learning procedures for a better adjustment of the thresholds (t1 and t2) and a more precise definition of the situations which activate the search for best solutions, the last step of the process. What was learned when running the system on contemporary documents is the need to improve graphics segmentation to avoid the confusion of small graphic components with text. The system was designed to provide assistance to the archivists when extracting the FC-pairs and to enable a semi-automatic procedure to recover from errors. The design of an interactive tool is foreseen.

8. REFERENCES

[1] Aiello, M. and Smeulders, A. M., "Thick 2D relations for document understanding," Information Sciences-Informatics and Computer Science 167 (1-4), Elsevier Science, New-York, 147-176 (2004).

[2] Faure, C., Vincent N., “Simultaneous detection of vertical and horizontal text lines based on perceptual organization”. Proc. SPIE 7247, Document Recognition and Retrieval XVI, San Jose, CA, USA (2009).

[3] Gupta, S., Kim, J., Grauman, K. and Mooney, R.J., "Watch, Listen & Learn: Co-training on Captioned Images and Videos," Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2008), Antwerp, Belgium, 457-472 (2008).

[4] Harrington, S. J., Naveda, J. F., Jones, R. P., Roetling, P., and Thakkar, N., "Aesthetic measures for automated document layout," Proc. Symposium on Document Engineering (DocEng '04), Milwaukee, Wisconsin, USA, ACM Press, 109-111 (2004).

[5] Khurshid, K., Faure, C., Vincent, N., “Fusion of Word Spotting and Spatial Information for Figure Caption Retrieval in Historical Document Image,” Inter. Conference on Document Analysis and Recognition (ICDAR), Barcelona, Spain, 266-270, (2009).

[6] Khurshid, K., Siddiqi, I., Faure, C., Vincent, N., "Comparison of Niblack inspired binarization techniques for ancient document images," Proc. SPIE 7247, Document Recognition and Retrieval XVI, San Jose, CA, USA (2009).

[7] Li, L., Lu, S., and Tan. C.L., "A graphic Image processing system," Proc. 8th IAPR International Workshop on Document Analysis System (DAS 08), IEEE, 455-462 (2008).

[8] Maderlechner, G., Panyr, J., Suda, P. "Finding Captions in PDF-Documents for Semantic Annotations of Images.," Lecture Notes in Computer Science 4109, Springer-Verlag, 422-430 (2006).

[9] Mao, S., Rosenfeld, A. and Kanungo, T.,"Document Structure Analysis Algorithms: A Literature Survey," Proc. SPIE 5010, Document Recognition and Retrieval X, San Jose, CA, USA, 197-207, (2003).
[10] Marshall, C.C., Shipman, F.M. III. "Spatial Hypertext: Designing for Change," Communications of the ACM 38 (8), 88-97 (1995).
[11] Quattoni, A., Collins, M., Trevor Darrell, T., "Learning Visual Representations using Images with Captions." Proc. Computer Vision and Pattern Recognition, CVPR 2007, IEEE CS Press, (2007).
[12] Srihari, R.K., "Use of Captions and Other Collateral Text in Understanding Photographs," Artificial Intelligence Review 8, Kluwer Academic Publishers, 409-430 (1995).
[13] BIUM: Bibliothèque InterUniversitaire Médicale, Paris, http://www.bium.univ-paris5.fr/histmed/medica.htm