Layout analysis on newspaper archives

Vincent Buntinx
vincent.buntinx@epfl.ch
École Polytechnique Fédérale de Lausanne,
Switzerland

Frédéric Kaplan
frederic.kaplan@epfl.ch
École Polytechnique Fédérale de Lausanne,
Switzerland

Aris Xanthos
aris.xanthos@unil.ch
Université de Lausanne, Switzerland

The study of newspaper layout evolution through historical corpora has been addressed by diverse qualitative and quantitative methods in the past few years (Antonacopoulos et al, 2013; Gonzalez et al, 2001; Liu et al, 2001; Mitchell and Hong, 2004; Singh and Bhupendra, 2014). The recent availability of large corpora of newspapers is now making the quantitative analysis of layout evolution ever more popular. This research investigates a method for the automatic detection of layout evolution on scanned images with a factorial analysis approach. The notion of eigenpages is defined by analogy with eigenfaces used in face recognition processes. The corpus of scanned newspapers that was used contains 4 million press articles, covering about 200 years of archives. This method can automatically detect layout changes of a given newspaper over time, rebuilding a part of its past publishing strategy and retracing major changes in its history in terms of layout. Besides these advantages, it also makes it possible to compare several newspapers at the same time and therefore to compare the layout changes of multiple newspapers based only on scans of their issues.

Introduction to the Corpus

The corpus consists of digitized facsimiles of two Swiss newspapers, "Journal de Genève" (JDG) from years 1826 to 1997 and "Gazette de Lausanne" (GDL) from years 1804 to 1997. Scanned daily issues of each journal were transcribed using an optical character recognition (OCR) system (Rochat et al, 2016). The entire scanned data weighs more than 20TB, which makes most usual analysis techniques out of reach for regular desktop computers. This corpus has been the focus of several studies analyzing textual data (such as linguistic changes (Buntix et al, 2016) and named entity recognition (Ehrmann et al, 2016)). An example of different layouts of GDL's first page is given in Figure 1 which shows the evolution of various features, such as title size and position, fonts and number of columns.

Bitmap Factorial Analysis

In order to analyze layout evolution, we propose to build a static layout representation for every year in the corpus. Thus, when studying each newspaper's first page, we define the pixel t of the static representation $\overline{P_{y,m}}$ of month m of year y as

$$\overline{P^t_{y,m}} = \frac{1}{N_{ym}} \sum_{d=1}^{N_{ym}} P^t_{y,m,d}$$

Where N_{ym} is the number of issues in month m of year y and $P_{y,m,d}$ is the first page of day d of month m of year y. The pixel t of the static representation $\overline{\overline{P_y}}$ of year y is then defined as

$$\overline{\overline{P^t_y}} = \frac{1}{N_y} \sum_{m=1}^{N_y} \overline{P^t_{y,m}} = \frac{1}{N_y} \sum_{m=1}^{N_y} \frac{1}{N_{ym}} \sum_{d=1}^{N_{ym}} P^t_{y,m,d}$$

Were N_y is the number of month representations $\overline{P_{y,m}}$ in year y.

A diagram of the process is shown in figure 2.

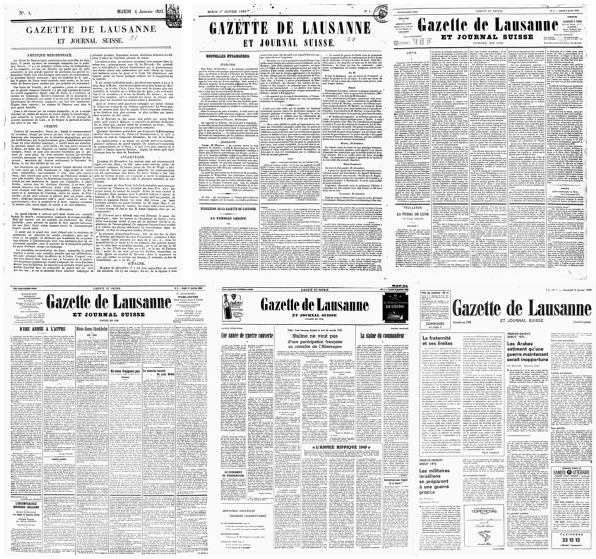


Figure 1. Different layouts of GDL in years 1825, 1850 and 1875 (top, left to right), 1925, 1950 and 1975 (bottom, left to right).

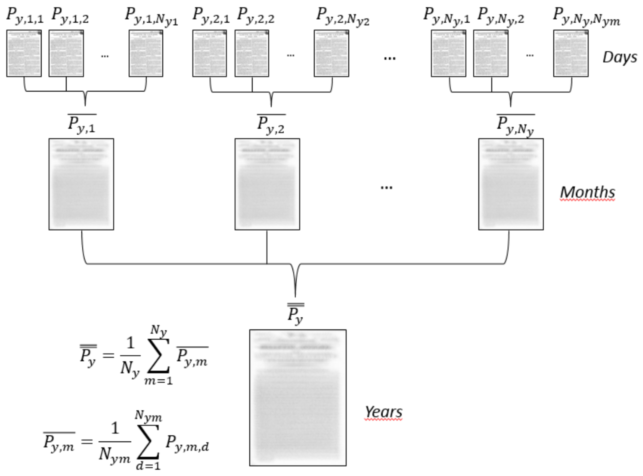


Figure 2. Process diagram creating a yearly representation of first page layouts.

These representations give a vision of the mean layout over the course of a given year. Each yearly representation can be projected in a two-dimensional space by performing a principal component analysis (PCA) which maximizes the covariance on every pixel. This method is analogous to the eigenfaces method used for face recognition (Turk and Pentland, 1991a, 1991b) We compute the eigenvectors, that we named eigenpages, as well as the eigenvalues of the covariance matrix of the pixels. The yearly representations are then projected in the two-dimensional space of the two eigenvectors which have the highest eigenvalues. The resulting projections of yearly mean images of JDG and GDL from years 1900 to 1998 are portrayed in Figure 3. In these figures, each point is a yearly image and consecutive years are linked in order to highlight the change over time. The further apart the points are, the bigger the layout's changes occurring between two years. Visual inspection reveals several clusters of years with a similar layout. Furthermore, homogeneous sequences of years may be clustered automatically based on the (unprojected) distance between them (e.g. by computing the distance between year y and y+1 and "cutting" the sequence of years at positions where their distance exceeds an arbitrary threshold.

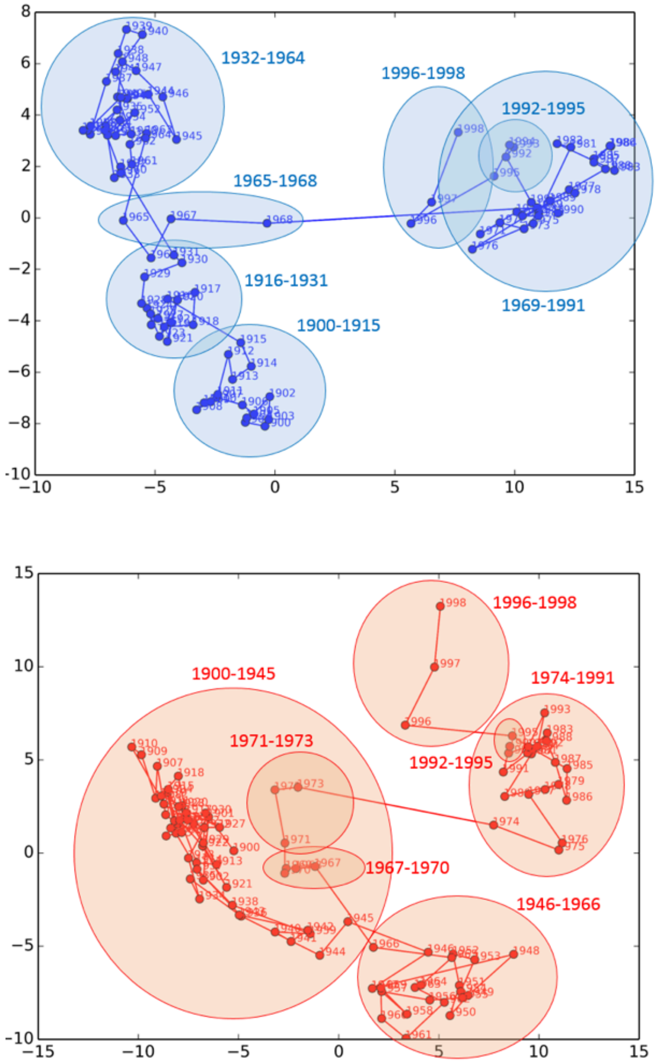


Figure 3: PCA projected results of the yearly representations of first pages of JDG (top, blue) and GDL (bottom, red) from years 1900 to 1998 with clusters obtained by visual inspection.

Discussion

The PCA technique allows us to quantify layout changes by covariance analysis of the pixels of yearly representations. The proportion of covariance information shown by the PCA is 73% for JDG and 76% for GDL. Visual interpretation reveals different chronological clusters which are displayed in Tables 1 and 2 along with their mean positions in the two-dimensional space of eigenpages as well as mean images representing these periods (computed in the same way as yearly images, cf. Figure 2). These mean images reveal the major layout transitions in each journal which may be summarized as follows:

Journal de Genève (JDG):

- 1900-1915: 6 columns, title above columns 2 to 5, little space between columns.
- 1916-1931: 4 columns, title above columns 1 to 4, more space between columns.
- 1932-1964: 4 columns, change of the layout around the title and the first title position.
- 1965-1968: 4 columns, change of the layout around the title, boxes with black borders begin to appear.
- 1969-1991: 4 columns, total change of the title, title above columns 2 to 4, logo appears, more space between columns and boxes, article titles are bigger.
- 1992-1995: 5 columns, fusion of JDG and GDL, big change of layout, boxes inside boxes begin to appear, more stable structure.
- 1996-1998: 6 columns, big change in title font, previous column layout replaced by a more classic one, article titles are placed at the top of the first page.

Gazette de Lausanne (GDL):

- 1900-1945: 6 columns, title above columns 2 to 5, little space between columns.
- 1946-1966: 7 columns, title above columns 2 to 6, more space between columns yielding particularly small column sizes.
- 1967-1970: 5 columns, title above columns 2 to 5, first column begins before the title which is on the right, advertisements placed below the page.
- 1971-1973: 6 columns, more classic layout with article titles at the top.
- 1974-1991: 4 columns, lots of space between columns and articles, bigger article titles.
- 1992-1995: 5 columns, fusion of JDG and GDL, big change of layout, boxes inside boxes begin to appear, more stable structure.
- 1996-1998: 6 columns, big change in title font, column layout replaced by a more classic one, the article titles are placed at the top of the first page.

The automatic clustering method described in previous chapter has been applied on unprojected distances and produce similar clustering results (depending on the threshold parameter). Qualitative analysis confirms that the resulting clusters are all separated by important layout transition phases.

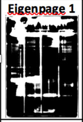
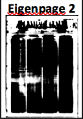
Journal de Genève	1900 – 1915 	1916 – 1931 	1932 – 1964 	1965 – 1968 	1969 – 1991 	1992 – 1995 	1996 – 1998
Eigenpage 1 	-1.5858	-4.5761	-6.3541	-4.0387	11.5680	9.6320	6.6127
Eigenpage 2 	-6.9613	-3.4511	4.0575	-0.4756	0.9258	2.3969	1.2411

Table 1: Chronological clusters with their mean first page representations and their positions in the axes of PCA eigenpages (JDG). PCAPCgenpag(JDG)obtained by PCA for JDG.

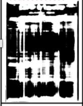
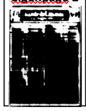
Gazette de Lausanne	1900 – 1945 	1946 – 1966 	1967 – 1970 	1971 – 1973 	1974 – 1991 	1992 – 1995 	1996 – 1998
Eigenpage 1 	-6.7016	4.5705	-2.1566	-2.6345	9.8505	9.5849	4.3915
Eigenpage 2 	0.4252	-7.1833	-0.8661	2.4913	3.9961	6.3149	10.0302

Table 2: Chronological clusters with their mean first page representations and their positions in the axes of PCA eigenpages (GDL).

This analysis is also useful to compare several newspaper publishing strategies. We projected the two newspapers in the same two-dimensional space representation (presented in Figure 3) using the same method with yearly representations of both journals in order to compare their chronological trajectories. The covariance information shown by the PCA is 67%. Visual inspection reveals three main clusters for each journal. Each of these clusters turns out to correspond to groups of clusters that has been detected in the previous projections. We observe that the layout of both journals has evolved in a similar way but with different timescales. GDL is more dispersed than JDG and has explored different strategies during the period 1900-1966. However, GDL has adopted a style more similar to JDG style between 1967 and 1973 just before it entered a major layout transition in 1974 (5 years later than JDG).

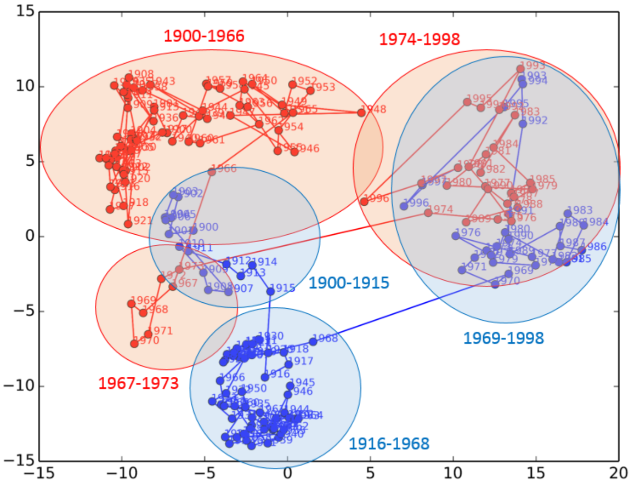


Figure 4. PCA projected results of the yearly representations of first pages of JDG (blue) and GDL (red) from years 1900 to 1998 in the same two-dimensional space representation with clusters obtained by visual inspection.

Conclusion

These first results demonstrate a promising method of detecting layout evolution automatically. The method is applicable to a large variety of longitudinal image corpora without any prerequisites, since it only requires images in bitmap format. It make it possible to compare several corpora and determine periods of layout transitions in a common two-dimensional space for visual interpretation. In addition, unprojected distances can be used to determine layout changes in an entirely automatic fashion, by analyzing the representation space through clustering algorithms. Future work on this method should include the integration of an alignment method in the bitmap preprocessing step, because alignment errors may impact the pixel covariance analysis and eigenpages creation.

Bibliography

Antonacopoulos, A., Clausner, C., Papadopoulos, C., and Pletschacher, S. (2013) ICDAR2013 Competition on Historical Newspaper Layout Analysis – HNLA2013. *12th International Conference on Document Analysis and Recognition*.

Buntinx, V., Bornet, C., and Kaplan, F. (2016) Studying Linguistic Changes on 200 Years of Newspapers. 2016. *DH2016*, Kraków, Poland, July 11-16.

Ehrmann, M., Colavizza, G., Rochat, Y., and Kaplan, F. (2016). Diachronic Evaluation of NER Systems on Old Newspapers. *13th Conference on Natural Language Processing (KONVENS 2016)*, Bochum, Germany, September 19-21.

González, J., Rojas, I., Pomares, H., Salmerón, M., Prieto, A., and Merelo, J.J. (2001) Optimization of web newspaper layout in real time. *Computer Networks*, Volume 36, Issues 2–3, July, Pages 311-321, ISSN 1389-1286, http://dx.doi.org/10.1016/S1389-1286(01)00158-X.

Liu, F., Luo, Y., Yoshikawaf, M., and Dongcheng, H. (2001). A New Component based Algorithm for Newspaper Layout Analysis. 2001. *6th International Conference on Document Analysis and Recognition*.

Mitchell, P. E., and Hong, Y. (2004) Newspaper layout analysis incorporating connected component separation. *Image and Vision Computing*, Volume 22, Issue 4, 1 April, Pages 307-317, ISSN 0262-8856, http://dx.doi.org/10.1016/j.imavis.2003.11.001.

Rochat, Y., Ehrmann, M., Buntinx, V., Bornet, C., and Kaplan, F. (2016). Navigating through 200 years of historical newspapers. 2016. *iPRES*, Bern, October 3-6.

Singh, V., and Bhupendra, K. (2014). Document layout analysis for Indian newspapers using contour based symbiotic approach. 2014. *International Conference on Computer Communication and Informatics (ICCCI-2014)*, Jan. 03 – 05, Coimbatore, INDIA

Turk. M., and Pentland, A. (1991a) Face recognition using eigenfaces. 1991. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. pp. 586–591.

Turk. M., and Pentland, A. (1991b) Eigenfaces for recognition. *Journal of Cognitive Neuroscience*. 3 (1): 71–86. doi:10.1162/jocn.1991.3.1.71. PMID 23964806.