



HAL
open science

Newspaper Magnification with Preserved Entry Points

Sebastian Gallardo, María Cristina Riff, Dorian Mazauric, Pierre Kornprobst

► **To cite this version:**

Sebastian Gallardo, María Cristina Riff, Dorian Mazauric, Pierre Kornprobst. Newspaper Magnification with Preserved Entry Points. 2023. hal-04210840

HAL Id: hal-04210840

<https://hal.science/hal-04210840>

Preprint submitted on 19 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Newspaper Magnification with Preserved Entry Points

Sebastian Gallardo^{a,b,c}, María Cristina Riff^c, Dorian Mazauroic^a, Pierre Kornprobst^a

^a*Université Côte d’Azur, Inria, France*

^b*Demain Un Autre Jour, France*

^c*Universidad Técnica Federico Santa María, Chile*

Abstract

We present a new task called “Newspaper magnification with preserved entry points”, which consists in making the print edition more readable and usable in small displays by preserving the quality of entry points (e.g., headlines, images, illustrations), i.e., all elements that allow to scan a page and focus in the article they want. Pinch-to-zoom is the classical feature available to magnify the content. However, the pinch-to-zoom feature has a well-known drawback: one quickly loses positioning inside the page and thus can not scan it anymore, significantly impacting reading comfort and navigation. To solve this problem, we propose a formulation that preserves entry points during magnification while keeping all stories on a page always visually present. Our approach proposes a new packing problem to generate alternative layouts considering computational aesthetics criteria, which we solve by a genetic algorithm. As a proof-of-concept, we focus here on text-only content and headlines readability and usability in this paper. Results show that our approach improves entry point quality in the generated layouts compared

Email address: sebastian.gallardo-diaz@inria.fr (Sebastian Gallardo)

to a simpler magnification using the original layouts. We also show how aesthetic criteria enable better article placements (modifying the aspect ratio of some articles) without impacting the overall headline quality. We believe the Newspaper magnification with preserved entry points task will motivate the community along the line of research for document engineering, user interface, and accessibility.

1. Introduction

While there have been a lot of studies and debates around print versus online newspapers, print newspapers continue to be a point of reference in terms of quality and user experience ([8]). This attachment we have for print can be explained mainly by the beauty of their design, which provides a strong, consistent, and appealing visual approach to enhance a story's impact and comprehension ([10, 7]), together with better recognition of newspaper branding. However, combining the design and aesthetics from the printed edition and the functionalities of the online edition remains a complex design challenge ([17, 15]). In this sense, we need to go beyond the limits of the two main current modes of access to the digital press, which are: (1) access to an electronic version of the printed one, on a reader or tablet, through digital kiosk applications (access to more or less enhanced PDF files, in general); (2) access to a web version, adaptable to various screen sizes thanks to "responsive design", which allows us to read articles but misses out on the advantages of printed newspapers.

Print newspaper reading experience is indeed very particular. While readers face a newspaper page, numerous behavioral studies have shown that the

page is not read linearly and exhaustively. In [9], the authors have highlighted the existence of *entry points* (e.g., photos, illustrations, headlines) that provide readers with the necessary elements to scan newspapers and select what they want to read. As such, these entry points are essential in the print user experience.

Because of their importance, newspaper designers and journalists follow some guidelines and good practices to make entry points efficient. For example, for a headline to work as intended, the font plays a crucial role, but one must also consider the number of lines to split the headline string into. Indeed, most newspaper headlines occupy one or two lines (sometimes three or four, but never more). This empiric choice is probably related to the fact that return sweep saccades (i.e., saccades to move from the end of a line to the beginning of the next one) substantially impact reading time ([14]) and thus reading comfort. In fact, a study made by [25] shows that return sweep saccades are heavily impacted by the font size and line size, both in the position of landing in the next line and the probability of an under-sweep (a return sweep that lands too far away from the left margin and needs a correction). Also, other works study general assumptions about entry points with a socio-semiotic perspective together with experiments using eye-tracking technologies to confirm these assumptions ([16]).

Since online newspapers have been fundamentally a simple transposition of the print newspaper into digital media, one might think that the print experience should be the same. Unfortunately, this simple reasoning does not work simply because print newspapers are not designed to be read on small displays, so one cannot expect that scanning will be as pleasant and efficient

as in a smaller-scale version. To convince yourself, let us consider here the case of headlines, which will be our focus in this paper, and the example shown in Fig. 1(a) where the headlines are too small and not *readable* so that they can not be used to scan the page and choose the story to read.



(a) Input (b) Pinch-to-zoom (c) In-place mag. (d) NMPEP

Figure 1: Principle of Newspaper magnification with preserved entry points (NMPEP) with a magnification factor of two: (a) Example of print newspaper page where each article has been colored to highlight the structure. On small displays, headlines become too small to be readable and usable to scan the page. Magnification is needed. (b) Pinch-to-zoom method induces the common local/global navigation difficulty encountered when reading newspapers via digital kiosk applications. Content is magnified, but the global view of the page is lost. (c) In-place magnification, i.e., one magnifies the content while keeping the original layout. In that case, headlines become more readable (because of the font size increase) but less usable (because they are split into too many lines here five articles have more than 3 lines). (d) Solution targeted by NMPEP, where a new layout is proposed so that all articles are shown, and headlines are readable but also more usable (only one article has more than three lines - compare to (c)).

Thus, obviously, to see headlines better, one needs to magnify them. On

touch screens, pinch-to-zoom is the classical feature available to magnify content. As illustrated in Fig. 1(a)), the pinch-to-zoom feature has a well-known drawback: one quickly loses positioning inside the page and thus can not scan it anymore, significantly impacting reading comfort and navigation. Another method could be to increase the font size while keeping the layout fixed, as shown in Fig. 1(c). We call this method *in-place magnification*. With that method, most headlines will naturally flow on more lines and thus become less *usable*, i.e., working poorly as a good hook. Note that, of course, with that solution, since we simply increase fonts while keeping the same area of the display, some text will not fit anymore in the display, which is not a problem since we target digital media. What concerns us here is to enable scanning and selection of a story, and when a story is selected, one could switch to a single-page mode to read the story selected.

To address these problems, we propose a new task called “entry point preserving newspaper magnification”, which makes the print edition readable and usable in small displays by preserving the quality of entry points. As proof-of-concept, we focus here on text-only content and headlines as entry points. The general idea is to explore alternative layouts where we can change the aspect ratios of articles so that headlines will be displayed in large fonts and a few lines. The difficulty is that changing the aspect ratio of one article induces a domino effect requiring a complete re-layout of the page by considering alternative shapes for all articles (even if they were not *a priori* impacted by the magnification). An example of an alternative layout that we want to generate is shown in Fig. 1(d). Headlines are not only more readable (referring to the font size) but also usable (referring to the number of lines

to split the headline string into).

To solve this task, we present an original approach formalized as a combinatorial/geometric packing problem. Our input is an existing layout and a magnification factor. Given each block of text, a finite set of admissible shapes can be inferred, given constraints related to aesthetics and print size; then, one needs to pack these shapes – a Tetris-like problem where one shape per block must be selected to create a new layout. Our formulation of the problem as a packing problem leads to an NP-hard optimization problem.

This paper is organized as follows. In Sec. 2, we briefly review the state-of-the-art in different topics related to our problem. In Sec. 3, we make a formal definition of our problem. In Sec. 4, we introduce our evolutionary approach called Evolutionary Newspaper Magnification (ENM) to tackle this problem. In Sec. 5, we show results on realistic examples to demonstrate the advantages of the proposed algorithm. Finally, in Sec. 6, we present the conclusions and insight for future improvement.

2. Related Work

Relayouting. Among methods for relayouting in newspaper research, it is interesting to notice the frequent use of genetic algorithms (GA) ([3, 23, 11]). The reason is that GA can incorporate soft constraints (aesthetic) in chromosomes and/or fitness function and hard constraints (overlapping, margin, etc.) or incorporate aspects other than position like fonts and content variation inside chromosomes. Also, in general, GA do not need a lot of specific low-level constraints, making the creation process simpler, and seems appropriate for such artistic tasks since, unlike other more brute force algorithms,

GA do not attempt to mimic or model any particular process by which solutions are created. Instead, solutions are generated randomly and evaluated after the fact. On the other side, systematic/exact approaches might be too large to solve real-life efficiency requirements.

Packing problems. *2D Packing Problems* are combinatorial optimization problems with a lot of applications, like cutting of raw-material or item packing. In all applications, the focus is the positioning of figures with known dimensions inside bigger containers. Some authors proposed exact methods using *Branch and Bound* or Dynamic Programming techniques to solve this kind of problem. However, in many instances, it is focused on *perfect packing* instances (with no empty spaces) and a low number of pieces (less than 30), and still obtaining long execution times. Given that issue, heuristic methods are the most-used tool to solve it, like packing pieces using Bottom Left strategy or in the space that minimizes wasted space (Best Fit). In the survey made by [22], different improvement heuristics are classified in search over sequences and direct search over the layout itself. Heuristics that search over sequences codify the layouts in some data structure (like an array), and perform modification operators like swapping or inserting to construct new layouts after decoding using a specific method ([4, 24, 5]). On the other hand, heuristics that search over layouts directly work with the pieces, like in [1].

Computational aesthetics. An important aspect of newspaper design is aesthetics, which is naturally subjective but can be measured considering specific criteria. Some classical aspects considered are balance, equilibrium, or symmetry. In [20], the authors propose 13 equations to measure different

aesthetic measures focusing on screen and web page design. Another related contribution is [12], where the authors propose measures for document layouts with a more algorithmic approach and slightly different definitions. We based our approach on these two works.

3. Problem Formalization

3.1. Notations

Let us consider a newspaper page composed of a set of N articles:

$$\mathcal{A} = \{\mathcal{A}_i\}_{i=1,\dots,N},$$

where we separate the content and the geometry clearly.

Concerning the content, in this paper, we assume that each article only contains text and can be described by:

$$\mathcal{A}_i = \{\mathcal{H}_i, \mathcal{B}_i\},$$

where \mathcal{H}_i is the headline and \mathcal{B}_i is the body text. The content is fixed; only the headline's length will matter in the sequel. For the sake of simplicity (but without loss of generality for our approach), we assume that all headlines share the same font size equal to α_0 .

Concerning the geometry, the newspaper page is defined by a layout \mathcal{L} , which is described by the position and size of each article \mathcal{A}_i composing it:

$$\mathcal{L} = \{\mathcal{L}_i\}_{i=1,\dots,N},$$

where

$$\mathcal{L}_i = \{x_i, y_i, w_i, h_i\},$$

where, for each article \mathcal{B}_i , (x_i, y_i) are the coordinates of the upper-left corner, w_i is the width and h_i is the height. So these values fully define the geometry of each article, and so the page. In the sequel, we will explore how to produce alternative layouts, i.e., layouts with the same article area distribution but arranged differently and with possible variations in $\{w_i, h_i\}$ values.

3.2. Decision and Optimization Problems

The input of our problem is a newspaper page defined by a set of articles $\mathcal{A} = \{\mathcal{A}_i\}_{i=1,\dots,N}$, and an original layout $\bar{\mathcal{L}} = \{\bar{\mathcal{L}}_i\}_{i=1,\dots,N}$.

Then let us first define the function λ :

$$\lambda : (\mathcal{H}_i, w_i, \alpha_0) \mapsto \text{number of lines headline } \mathcal{H}_i \text{ is split into.}$$

This function obviously depends on the width of the article (w_i) and the font size of the headline (α_0). The algorithm used for splitting the headline into several lines is presented in the appendix to be self-consistent but note that any other algorithm could be used instead.

Given these definitions, we can ask the following questions: Considering that a reader wants to increase the font size by a magnification factor of α , is the original layout "satisfactory", i.e., do we have that $\lambda(\mathcal{H}_i, \bar{w}_i, \alpha\alpha_0) \leq \lambda_{max}$ for all articles (where λ_{max} is a parameter)? If the answer is no, which is very likely since we increase the font size, can we find an alternative layout \mathcal{L}' where $\lambda(\mathcal{H}_i, w'_i, \alpha\alpha_0) \leq \lambda_{max}$ for all articles or at least many of them? And finally, can we find layouts that also follow some aesthetic criteria?

These questions can be translated formally into the following decision problem:

DECISION PROBLEM

Input:

A set of articles $\mathcal{A} = \{\mathcal{A}_i\}_{i=1,\dots,N}$, where $\mathcal{A}_i = \{\mathcal{H}_i, \mathcal{B}_i\}$, an initial layout $\bar{\mathcal{L}} = \{\bar{\mathcal{L}}_i\}_{i=1,\dots,N}$, where $\bar{\mathcal{L}}_i = \{\bar{x}_i, \bar{y}_i, \bar{w}_i, \bar{h}_i\}$, a headline font size α_0 , a magnification factor α , a threshold for the number of lines λ_{max} , $k_1 \geq 0$, $k_2 \geq 0$.

Question:

Given a function H that promotes headlines to be displayed in less than λ_{max} lines, and a function A that promotes aesthetics of layouts, does it exist an alternative layout \mathcal{L}' such that $H(\mathcal{A}, \mathcal{L}') \geq k_1$ and $A(\mathcal{L}') \geq k_2$?

Based on this decision problem, we propose the following optimization problem:

OPTIMIZATION PROBLEM

Input:

A set of articles $\mathcal{A} = \{\mathcal{A}_i\}_{i=1,\dots,N}$, where $\mathcal{A}_i = \{\mathcal{H}_i, \mathcal{B}_i\}$, an initial layout $\bar{\mathcal{L}} = \{\bar{\mathcal{L}}_i\}_{i=1,\dots,N}$, where $\bar{\mathcal{L}}_i = \{\bar{x}_i, \bar{y}_i, \bar{w}_i, \bar{h}_i\}$, a headline font size α_0 , a magnification factor α , a threshold for number of lines λ_{max} , $w_A \in [0, 1]$

Question:

Find the layout \mathcal{L}' solution of the problem:

$$\mathcal{L}' = \underset{\mathcal{L}}{\operatorname{argmax}} E(\mathcal{L}), \quad \text{with} \quad E(\mathcal{L}) = H(\mathcal{A}, \mathcal{L}) + w_A A(\mathcal{L}). \quad (1)$$

Now that the optimization problem is posed, we define the energy E in the next section.

3.3. Definition of the Energy (E)

3.3.1. Definition of the Headline Term (H)

We define $H(\mathcal{L})$ as:

$$H(\mathcal{A}, \mathcal{L}_{\mathcal{A}}) = e^{-\mu f(\mathcal{A}, \mathcal{L}_{\mathcal{A}})}, \quad (2)$$

where μ is a parameter and

$$f(\mathcal{A}, \mathcal{L}_{\mathcal{A}}) = \frac{1}{N} \sum_{i=1}^N \max \{ \lambda(\mathcal{H}_i, w'_i, \alpha\alpha_0) - \lambda_{max}, 0 \}^2.$$

Using this function, if we call excess the difference $\lambda(\mathcal{H}_i, w'_i, \alpha\alpha_0) - \lambda_{max}$ when it is positive, this term H means that high values of excess will be strongly penalized. When there is no excess, i.e., when $\lambda(\mathcal{H}_i, w'_i, \alpha\alpha_0) - \lambda_{max}$ is negative, it does not contribute to increase f and thus does not change H .

3.3.2. Definition of the Aesthetic Term (A)

We define $A(\cdot)$ by selecting three of the criteria used by [12, 20] that are the most relevant for our problem, namely alignment, regularity, and balance:

$$A(\mathcal{L}) = A_1(\mathcal{L}) + A_2(\mathcal{L}) + A_3(\mathcal{L}),$$

where:

- $A_1(\mathcal{L})$ refers to aligning the left, right, top, and bottom edges together (i.e., how the left edges of all shapes are aligned?). We started from the measure proposed by [12], but we considered all edges in our case. In Fig. 2, we present different cases. Figure 2(a) has many articles aligned respecting left edges, something not present in Fig. 2(c). Also, Fig. 2(a) has three articles aligned respecting right edges, which is different from

Fig. 2(c). However, Fig. 2(c) has better *vertical* alignment (top and bottom) than Fig. 2(a).

- $A_2(\mathcal{L})$ refers to regularity. This term suggests that it is best to distribute articles regularly, close to a regular grid of blocks. For example, it is better if rows and columns of a table have relatively the same heights and widths. We used the approach proposed in [12].
- $A_3(\cdot)$ refers to balance, which can be defined as the distribution of optical weight on the newspaper page. Here, we took the same definition as in [20]. An example of well-balanced layout is given in Fig. 2(b)

For more details about the equations of these measures, please refer to the Appendix 7.

4. Evolutionary Approach: Evolutionary Newspaper Magnification (ENM)

We propose an evolutionary algorithm to produce new layouts based on our previous formulation, for many reasons:

1. It is fast to implement, and convexity is not required.
2. Genetic approaches, as explained in Section 2, allow us to introduce many characteristics of the desired layouts inside the representation in a clean and simple way.
3. Genetic algorithm does not attempt to mimic or model any particular process by which solutions are created. Instead, solutions are generated randomly and evaluated after the fact [11].



(a)

(b)

(c)

Figure 2: Aesthetic terms illustrations. In these three layouts, we can observe that: (a) has nice alignment (almost all the articles are aligned to the left of the page), but not so great balance (related to the vertical center axis); (b) has a high balance but at acceptable alignment. (c) is a good trade-off between these two aesthetic measures. All three examples share similar regularity values since they follow a horizontal grid. Notice that even if (b) is a good layout related to balance, it is not the best in terms of $H(\cdot)$.

We show the general structure in Algorithm 1, and then we proceed to explain each part of it.

4.1. Notion of Shape and Generation of Alternative Shapes

To create a new layout $\mathcal{L}_{\mathcal{A}}$, it is necessary to define $\{x_i, y_i, w_i, h_i\}$ for each article \mathcal{L}_i . In particular, we deal with the dimensions $\{w_i, h_i\}$ by **pre-computing** each possible pair $\{w_i, h_i\}$, so the algorithm can use a certain aspect ratio. We call each $\{w_i, h_i\}$ an *alternative shape*. The process is illustrated in Fig. 3. More precisely, given an article \mathcal{A}_i , we define M_i alternative shapes $\{\mathcal{S}_{i,j}\}_{j=1,\dots,M_i}$ having the same area as the initial one, and widths being multipliers of a multi-column layout size γ that we impose over the newspaper page (which is something that newspaper creators do), adjusting the height in a way that preserves the article’s area, and of course fitting into the page. Note that we will use $j = 1$ to refer to the original shape (i.e., $\mathcal{S}_{i,1} = \bar{\mathcal{S}}_i = \{\bar{w}_i, \bar{h}_i\}$). In Fig. 3, we show five alternative shapes for \mathcal{A}_6 (including the original one). Note that the initial position should not be considered to build alternative shapes (they should only fit the page). As a definition, a shape will be either *allowed* if $\lambda(\mathcal{H}_i, w_i, \alpha\alpha_0) \leq \lambda_{max}$ or *unwanted* otherwise.

4.2. Chromosome Representation

A chromosome encodes, in our case, a single generated layout. In practice, each chromosome contains two lists:

- (i) A *shape* list describing the shape index of each article. For example, $[3, 1, 2, 1]$ indicates that, in this layout, article 1 use shape 3 ($\mathcal{S}_{1,3}$), article 2 use shape 1 ($\mathcal{S}_{2,1}$), etc.

Algorithm 1 ENM Algorithm structure

Input: Newspaper page defined by a set of text-only articles (headline, body text, size, and coordinates).

Generate and classify alternative shapes for each article, as allowed or unwanted (see Sec. 4.1).

Generate M random chromosomes using Initial Population Algorithm. (see Sec. 4.4).

Generate each layout by packing shapes of each chromosome, in the order and using the heuristic given by the chromosome, and compute the corresponding score given by the evaluation function.

Save the best chromosome for the next generation.

for each one of the N generations **do**

while $M - 1$ slots are not filled **do**

 Select two parents using selection procedure (see Sec. 4.5), and apply crossover, with probability p_c (see Sec. 4.6).

 Select one parent using the selection procedure, and apply mutation for permutations, with probability p_m (see Sec. 4.7).

 Generate each new layout to compute the corresponding score of the newly generated chromosomes.

 Save the result for the next generation.

end while

end for

Output: Population of chromosomes where each one represents an alternative layout.

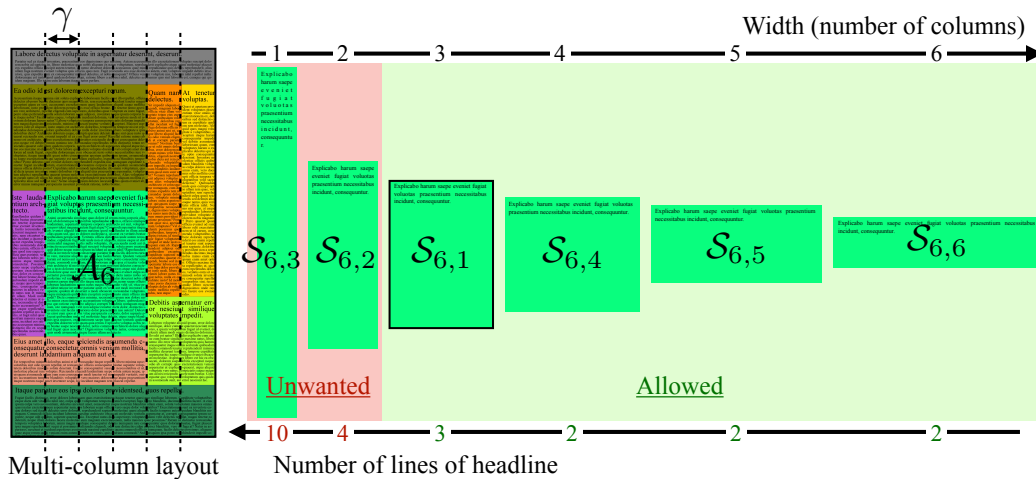


Figure 3: Generation of alternative shapes for \mathcal{A}_6 . Notice that each width is a multiplier of the column width γ . In that case, we have $\lambda_{max} = 3$ so that we have four allowed shapes and two unwanted shapes given the \mathcal{A}_6 headline.

- (ii) A *permutation* list that defines the order in which the articles are to be packed on the page following a fixed heuristic (see Sec. 4.3). For example, $[2, 1, 4, 3]$ indicates that the algorithm constructs the layout by first packing article 2, then article 1, etc.

Given a chromosome, one can render a newspaper page by applying the packing of the chosen article shapes and then filling in the shapes with the content (with the magnification factor applied to the text).

Using this type of representation, we can do simple swaps or similar operations over the chromosomes to search for alternative layouts. However, it can also represent a layout that is not realistic if it is impossible to pack the shapes in the given order without exceeding page dimensions. Thus, we say that a chromosome is *layout-compatible* if it can generate a layout given the dimensions of the page. The algorithm works with layout-compatible

chromosomes and discards any chromosome that does not have this property.

4.3. Heuristics

In this paper, we tested and compared three heuristics to pack shapes one by one, following the permutation list. These heuristics will be used to decode chromosomes into layouts. We chose them because they have been very studied in the packing problem literature. However, note that our genetic approach can use any kind of heuristic that packs pieces (articles in our case) without imposing their own order, i.e., places pieces one by one following the order given by a different data structure, which is, in our case, a permutation list. Our three heuristics are (see Fig. 4):

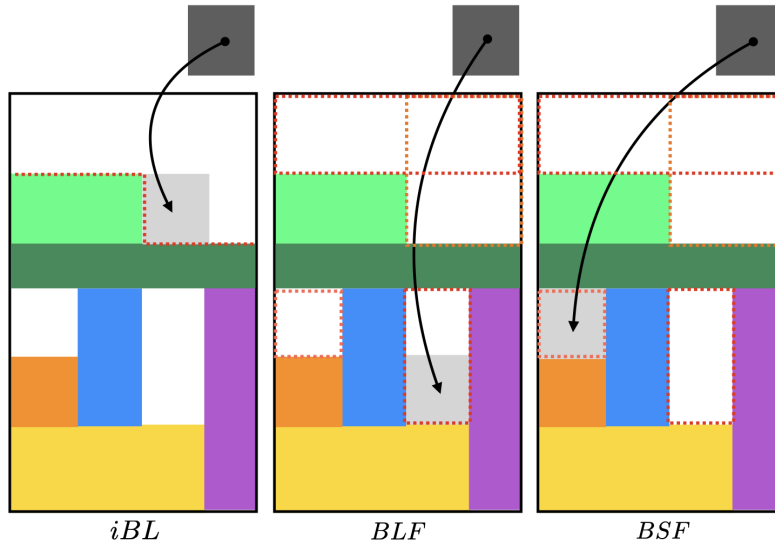


Figure 4: Illustration of the behavior of the three heuristic; starting from the same initial layout, we show how each heuristic places an article.

- *Improved Bottom Left* (iBL, see Fig. 4(a)): [2] introduced Bottom Left

heuristic, which tries to identify a space for a specific piece. Just as suggested by the name, it tries to place each piece starting from the top of the page and proceeds to move the piece to the bottom and then to the left, alternating between these moves until it cannot be moved anymore, placing the piece in the final position at the most possible bottom-left position of the page. The main advantage of the bottom-left strategy is being very fast (With complexity $O(n^2)$). In [19], the authors proposed *improved BL*, which supposes a BL-improvement, always gives priority to the bottom movement, i.e., when it is impossible to move the rectangle downwards, it moves it to the left, but just the distance necessary to make a downward movement again. To implement iBL, the algorithm keeps track of the “skyline” of the page and places each article in the bottom-left corner.

- *Bottom Left Fill* (BLF, see Fig. 4(b)): In [6], the approach searches just as BL but allows one to look in any space, even holes formed by pieces around. This increases the search space, leading to – in general – worse execution times, but allows to fill holes present in the page. The implementation of this heuristic was made by keeping track of all rectangular spaces, packing each article in the bottom-left space.
- *Best Space Fit* (BSF, Fig. 4(c)): We additionally implemented a variation of BLF that instead of selecting the bottom-left available position of the page, places each piece in the available space that minimizes the wasted space (i.e., the difference between space and piece areas).

Notice that each heuristic places articles in different positions based on

their own rules, leading to different layouts using the same permutation list. However, we can obtain the same layout by using different permutation lists.

4.4. Initial Population Algorithm

The goal is to generate an ensemble of M initial layout-compatible chromosomes. To do this, we first encode the original input as one possible solution. Then, we generate $M - 1$ random chromosomes iterating the following process: choose a random allowed shape for each article (pick an unwanted shape only if there is no allowed shape for an article) and a random permutation. If this chromosome is layout-compatible, we keep it, otherwise not. We iterate until we obtain the M chromosomes.

However, in practice, it may be difficult to find these layout-compatible chromosomes by a totally random procedure. So, we define a maximal number of iterations for this procedure, after which we relax the search of chromosomes.

When the algorithm is looking for a new chromosome, if we do not find any layout-compatible chromosome after T_1 iterations, the algorithm will allow the use of unwanted shapes. Finally, if we do not find a random solution after $T_2 > T_1$ iterations, it will repeat the original input as a solution and continue with the next chromosome.

4.5. Selection Procedure

A genetic algorithm picks, in each generation, one or two chromosomes to perform crossover or mutation, respectively. It chooses based on a specific called a *selection procedure*.

We propose a roulette wheel selection. This procedure selects one chromosome so that chromosomes with better score have more probability of being chosen. Here, the probability $p(\mathcal{L}_i)$ to be selected is computed as:

$$p(\mathcal{L}_i) = \frac{E(\mathcal{L}_i)}{\sum_{j=1}^M E(\mathcal{L}_j)}, \quad \text{for } i = 1, \dots, M.$$

We can see that our proposal is to use $E(\cdot)$ directly in the selection procedure.

4.6. Crossover Operator

The crossover operator works as follows:

- (i) It generates a first child that inherits the shapes list of the first parent and the permutation list of the second parent.
- (i) It generates a second child doing the opposite: shapes list of the second parent and permutation of the first one.

We keep the chromosomes if they are layout-compatible chromosomes. Otherwise, the first parent is selected for the next generation when the first child is not layout-compatible. Similarly, the second parent is selected for the next generation when the second child is not layout-compatible. We do it this way because we found it difficult to obtain a layout-compatible chromosome by only changing one individual shape. We found it more efficient to change the entire set.

4.7. Mutation for Permutations

To mutate the permutation list, we propose to generate two random numbers p, q , where $p \leq q$, and reverse the sublist composed by values with indices $i \in [p, q]$.

Again, only layout layout-compatible chromosomes are kept. If the child is not layout-compatible, its parent is selected for the new generation

4.8. Output

The output of the evolutionary algorithm is a population of chromosomes. Each one can be decoded to build a newspaper page.

5. Results

5.1. Implementation Details

Our method was implemented in a hybrid C++ programming language for ENM implementation and Python3 for extracting alternative shapes based on content. Tests were run in an Intel Core i7 machine of 4.2 GHz, using a Unix-based MacOS system. Unless stated differently, Table 1 shows parameter values used in our experiments.

5.2. Dataset

We built a dataset of 20 newspaper pages ($\{p_i\}_{i=1..20}$) whose layout was directly taken from real newspaper pages and changed the content to avoid any copyright issues. In a nutshell, (1) we used internal software to annotate real newspaper pages (here some *New York Times* front pages available online), essentially to decompose the page as a set of articles, (2) We filled each article with text-only content as assumed in that paper, with headline and body text from a *lorem-ipsu*m text generation. Some special care was taken to set the number of words in the headline to be realistic and match the original newspaper choices. By realistic, we mean that the number of words in the headlines was chosen so that the number of lines the headline

Parameter	Description	Value	Page
μ	Weight of $f(\cdot)$ function of headline term $H(\cdot)$	0.1	11
w_A	Weight of the aesthetic term	0.25	10
α	Magnification factor	2.0	9
γ	Multi-column grid width	387	14
λ_{max}	Maximum number of lines for headline	3	9
Q_1	Non-linear control parameter for $A_1(\cdot)$ term	100	42
Q_2	Non-linear control parameter for $A_2(\cdot)$ term	100	43
N	Number of generations	200	15
T_1	Iteration limit to use only allowed shapes in Initial Population Algorithm	250000	19
T_2	Iteration limit to search random chromosomes in Initial Population Algorithm	500000	19
M	Size of population	20	19
p_c	Probability to proceed with the genetic crossover operator	0.5	15
p_m	Probability to proceed with the genetic mutation operator	0.5	15

Table 1: Parameters chosen for our experiments.

split into (given the width of the article it belongs to) is consistent with what is observed statistically in the original newspaper. Figure 5 shows a sample of some pages from our dataset ¹.

5.3. Energy and Computational Time vs Heuristics

Respecting the experimental design, we run our algorithm testing for each one of the three heuristics. For each heuristic, we run the algorithm 20 times, each one with a different random seed. The idea is to keep the best solution in each run (best in the sense of higher energy). As a result, after 20 runs for each heuristic, we have 20 proposed layouts, with a total of 60 for each input considering all of the three heuristics.

¹Our dataset and visualization tools to produce pages as shown in the paper are available at <https://github.com/sgallard/ENMa-Instances-and-Tools>

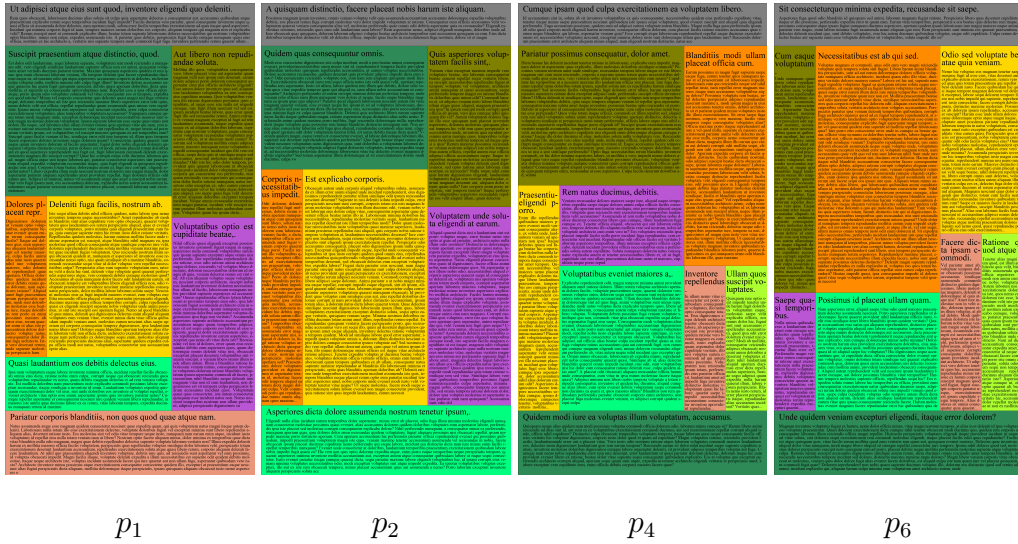


Figure 5: Sample pages from our dataset. The pages from our dataset were made by using layouts from real newspapers and filling in articles with text-only content.

5.3.1. Does Solutions Quality Depend on Heuristics?

In this section, we want to investigate if the quality of the $20 \text{ inputs} \times 60$ generated layouts = 1200 layouts depends on heuristics. Here, the quality of solutions is simply measured by their energy. First, we made some graphs and ran statistical analyses to test the quality of solutions found by the proposed method using different heuristics.

Boxplot in Fig. 6 and ANOVA test in Table 2 show no significant difference between heuristics and seeds. These results show that our method is stable, which has two consequences: (1) Using any heuristic to pack pieces is equivalent regarding the solutions' quality. This can be explained by using a permutation array to choose the packing order and the fact that using a permutation with a given heuristic is equivalent to using a different permutation with another heuristic. (2) Since there is no significant difference

between seeds, it is sufficient to run ENM only once with any seed, thus saving considerable time.

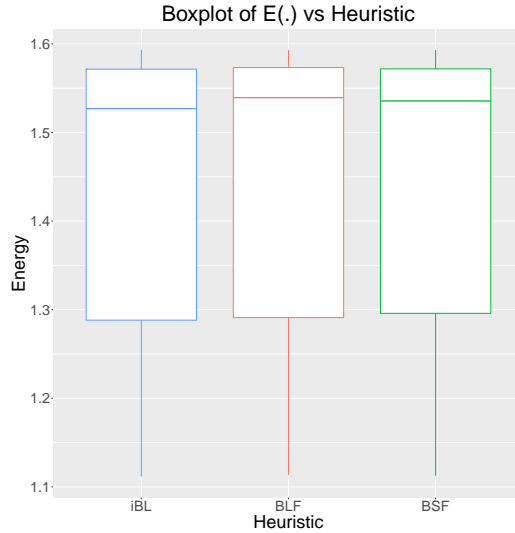


Figure 6: Boxplot comparing the energy with respect to heuristics. There is no statistical difference between heuristics.

5.3.2. Does Computational Time Depend on Heuristics?

We run a similar analysis for computational time, which is the total time (in seconds) that takes a single run of ENM. Figure 7 shows a difference between iBL and BLF/BSF heuristics, which we confirm using the ANOVA test shown in Table 3. Additionally, this test proves no significant differences in the computational time of the genetic approach for different runs with different seeds, which can be explained as related to the convergence of this kind of random procedure for this type of problem because there are geometrical and space constraints that make it converge. Then, we check pairwise comparisons between them. Results in Table 4 confirm a significant

Effect	DFn	DFd	F	p	p<.05
Seed	19.00	361.00	1.11	0.33	
Heuristic	1.02	19.45	1.24	0.28	
Seed:Heuristic	38.00	722.00	1.11	0.29	

Table 2: Two-way repeated measures ANOVA test for $E(\cdot)$. The columns in the table show the studied effect, numerator-degrees of freedom (DFn), denominator-degrees of freedom (DFd), F-statistic, and p-value. This shows a non-significant difference between random seeds and heuristics.

difference between iBL and the other two heuristics.

Finally, a more detailed analysis is shown in Fig. 8, divided for input layout. All examples show a significant difference, with BLF as the fastest heuristic (on average).

We found a significant difference in computational time, where BLF is the fastest, followed by BSF and finally iBL. This ranking can be explained regarding the implementation and characteristics of each heuristic. iBL is not capable of "filling holes" because the process only takes into account feasible movements across the available space. Because of that, in general, iBL needs more time to find an equivalent permutation list that allows packing a specific set of shapes, in comparison to the other two heuristics, which are capable of "filling holes" allowing finding optimal permutation lists faster. Regarding BSF, it can find equivalent results in a similar execution time to BLF, but it is a bit slower because it chooses the best shape, and BLF chooses the bottom-left free corner, something faster to compute.

Given results from Sections 5.3.1 and 5.3.2, from now all results discussed will be obtained using only the fastest heuristic (BLF).

Effect	DFn	DFd	F	p	p<.05	ges
Seed	4.74	90.13	0.44	0.81		0.00
Heuristic	1.03	19.57	9.58	0.01	*	0.02
Seed:Heuristic	38.00	722.00	1.23	0.16		0.00

Table 3: Two-way repeated measures ANOVA test for computational time (see columns meaning in the caption of Tab. 2). Results show a significant difference in computational time between heuristics but no statistical difference of seed and neither an interaction between heuristics and seeds.

	group 1	group 2	n_1	n_2	p.adj	p.adj.signif
Time	BSF	BLF	400	400	1.00	ns
Time	BSF	iBL	400	400	0.002	**
Time	BLF	iBL	400	400	0.00008	****

Table 4: Pairwise t-tests for computational time vs heuristics, with a Bonferroni adjust method for pairwise comparisons. Each row is a pairwise comparison between group 1 and group 2, each one with a number of examples n_1 and n_2 , respectively. It shows a significant difference between BSF and BLF vs iBL.

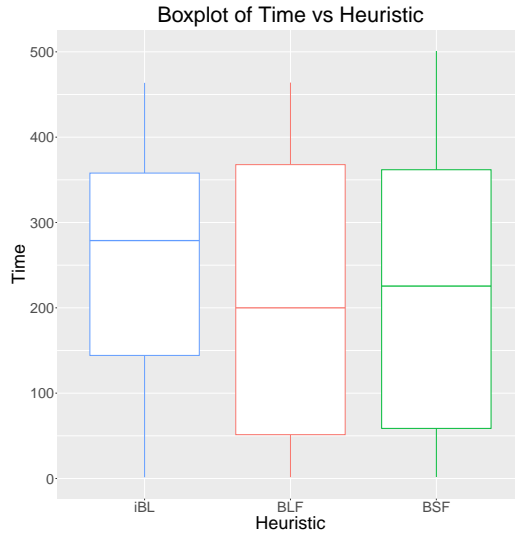


Figure 7: Boxplot comparing the computational time with respect to heuristics. There is a significant difference between iBL and the other two heuristics.

5.4. What is the w_A Impact in Each $H(\cdot)$ and $A(\cdot)$ Terms?

In this section, we compare the in-place magnification and our method, analyzing the headline term (H) and the aesthetic term (A) as separate terms. We used ENM with BLF heuristic and ran each configuration only once for each input layout. For a more complete analysis, we also discuss different values of w_A : (1) $w_A = 0$, to turn off the aesthetic term and see the impact in terms of the quality of the generated layouts; (2) $w_A = 0.25$, the standard configuration used in previous sections; (3) $w_A = 0.5$, to give more weight to the aesthetic term and see the impact in the distribution of articles.

Figure 9 shows that w_A does not have any impact on the value of H value, something not obvious *a priori* since giving more weight to the aesthetic term might change which shapes to use. Also, these graphs show a considerable

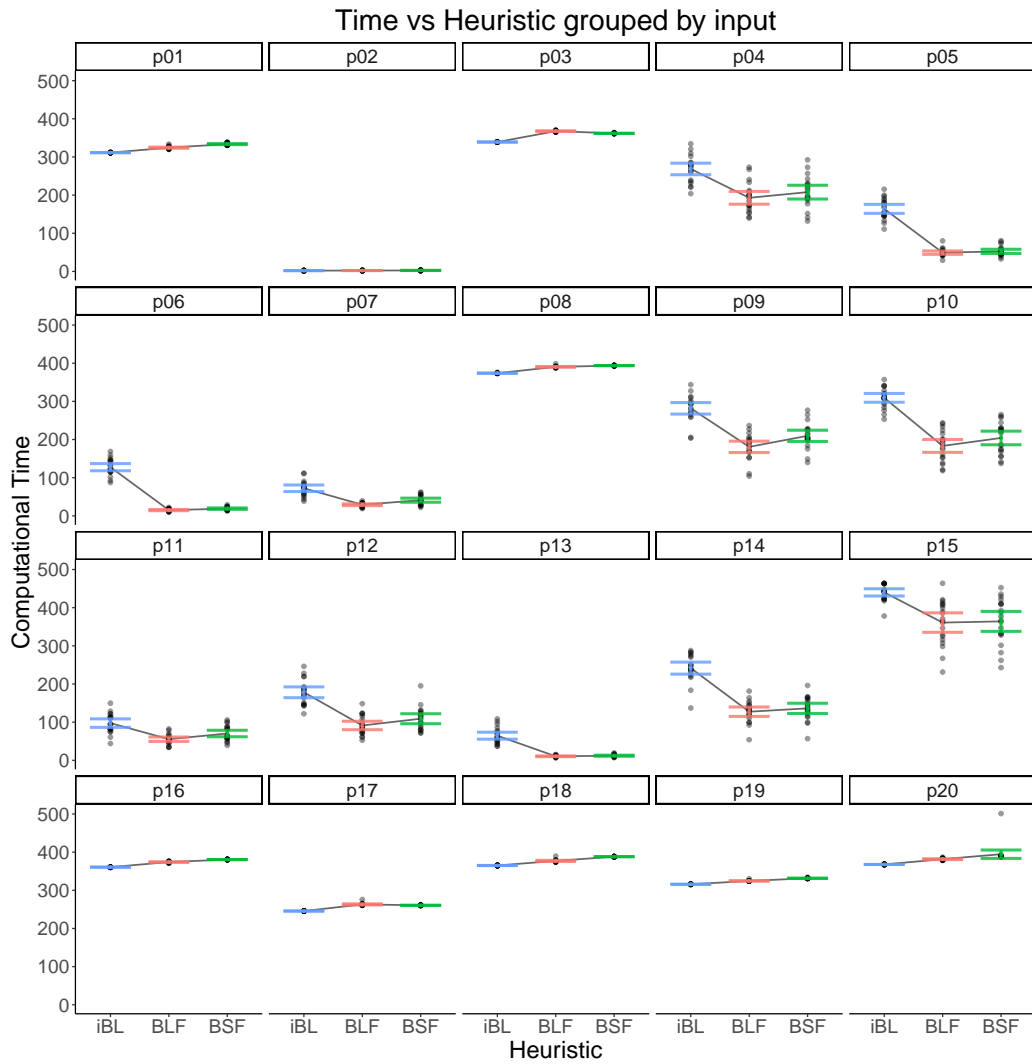


Figure 8: Comparison of the computational time depending on heuristics for each input. Each point corresponds to one execution for a given seed and heuristic and the corresponding computational time value. Means and 95% confidence intervals are shown in colors. Additionally, markers for each input's p-significance of ANOVA tests are shown at the bottom-left corner. We can see that, on average, BLF is the fastest, except in some cases where the intra-group deviation is very low (that can be seen by the range of the 95% confidence interval), where iBL is usually a bit faster.

increase in aesthetics between the in-place magnification and our method with $w_A > 0$. However, using $w_A = 0.5$ does not increase further aesthetic quality.

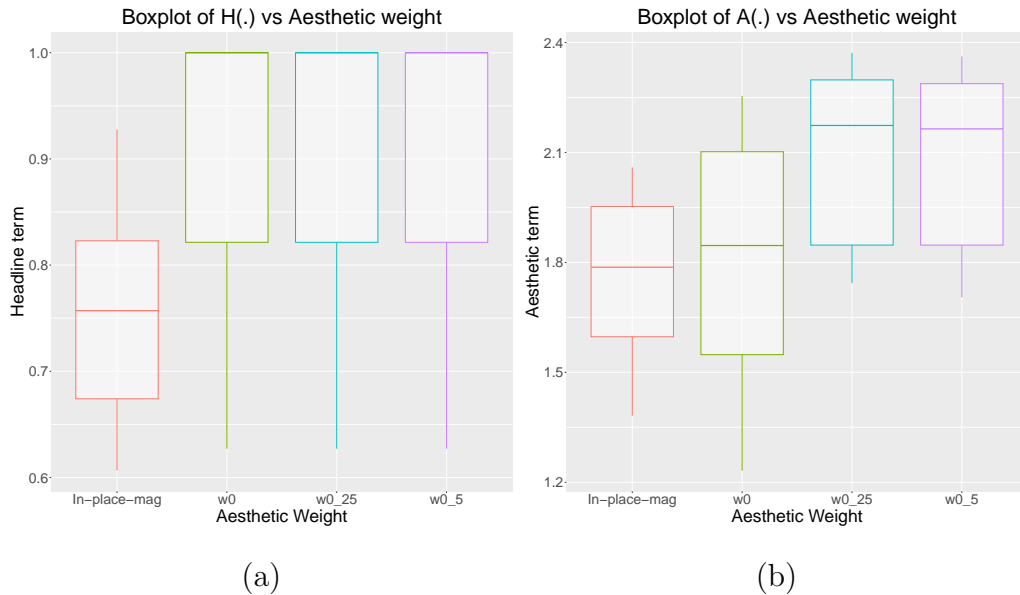


Figure 9: Comparison of the values of both terms in the energy (H and A different magnification methods). We show the results of the in-place magnification and our method using three different values for the weight w_A . (a) For the headline term, there is no difference between w_A values, and all of them show an improvement compared to the in-place magnification. (b) For the aesthetic term, there is an improvement when $w_A \geq 0.25$ compared to $w_A = 0$, but less difference between ENM with $w_A = 0.25$ and ENM with $w_A = 0.5$. Our method also gives better results in general than the in-place magnification.

To confirm these observations from Fig. 9, a non-parametric Friedman test of differences among repeated measures was conducted for this data and gave a Chi-square value of 36 for the headline term and 50.344 for the aesthetic term, both of them being significant ($p < .01$). These results indicate that there is a difference in both optimization terms between In-place mag. data

and different values of w_A .

To analyze this difference in more detail, in Tables 5 and 6, we ran pairwise comparisons between the in-place magnification and our method (with the different values of w_A). Results confirm that, for the aesthetic term, there is no difference between in-place magnification and our method with $w_A = 0$ (something that we expected). For our method, there is no difference between ENM with $w_A = 0.25$ and ENM with $w_A = 0.5$: since there are many geometrical constraints in this optimization problem, in addition to the possibility of using alternative shapes, the algorithm is not capable to improve aesthetics even more without impacting the overall $E(\cdot)$ value.

	group 1	group 2	n_1	n_2	p.adj	p.adj.signif
$H(\cdot)$	In-place mag.	$w_A = 0.0$	20	20	0.007	**
$H(\cdot)$	In-place mag.	$w_A = 0.25$	20	20	0.007	**
$H(\cdot)$	In-place mag.	$w_A = 0.5$	20	20	0.007	**
$H(\cdot)$	$w_A = 0.0$	$w_A = 0.25$	20	20	-	ns
$H(\cdot)$	$w_A = 0.0$	$w_A = 0.5$	20	20	-	ns
$H(\cdot)$	$w_A = 0.25$	$w_A = 0.5$	20	20	-	ns

Table 5: Pairwise t-tests for $H(\cdot)$ vs In-place mag. and ENM with different values of w_A , with a Bonferroni adjust method for pairwise comparisons (same presentation as in Tab. 4). It shows a significant difference between the In-place magnification and our approach but failed to reject the null hypothesis comparing different values of w_A , which we can interpret as no statistical difference given boxplots in Fig. 9

In Fig. 10(a), we see that in some cases, it is not possible to improve the headline term since the original layout is already near-optimal for a given magnification, but the method is capable of improving it in many cases. Also, in Fig. 10(b), we observe what we already discussed about $w_A = 0.25$

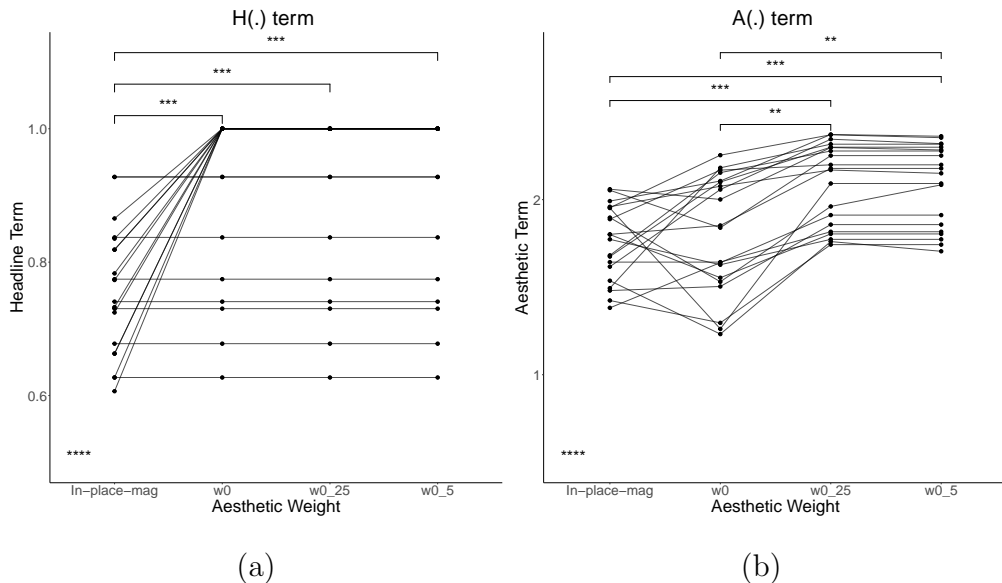


Figure 10: Values of $H(\cdot)$ and $A(\cdot)$ terms estimated from the solutions obtained by in-place magnification and our method (with different w_A values) for the 20 pages from our dataset. Each point corresponds to one result, and the lines connect the same input for different w_A values. Also, it shows when there is a significant difference between w_A values. (a) Results for the headline term ($H(\cdot)$) clearly show the improvement compared to the in-place magnification, but no difference between w_A values. Also, note that, in some cases, getting better results is impossible because of geometrical and/or alternative shape constraints. (b) Results for the aesthetic term ($A(\cdot)$) show an improvement compared to the in-place magnification. As expected, there is no improvement when $w_A = 0$. Additionally, it is interesting to see that there is no significant difference between $w_A = 0.25$ and $w_A = 0.5$, probably due to geometrical constraints that do not allow to improve aesthetics even more.

	group 1	group 2	n_1	n_2	p.adj	p.adj.signif
$A(.)$	In-place mag.	$w_A = 0.0$	20	20	0.54	ns
$A(.)$	In-place mag.	$w_A = 0.25$	20	20	0.0005	***
$A(.)$	In-place mag.	$w_A = 0.5$	20	20	0.0005	***
$A(.)$	$w_A = 0.0$	$w_A = 0.25$	20	20	0.0005	***
$A(.)$	$w_A = 0.0$	$w_A = 0.5$	20	20	0.0005	***
$A(.)$	$w_A = 0.25$	$w_A = 0.5$	20	20	0.54	ns

Table 6: Pairwise t-tests for $A(.)$ vs In-place mag. and different values of w_A , with a Bonferroni adjust method for pairwise comparisons (same presentation as in Tab. 4). It shows a significant difference between the in-place magnification and ENM with $w_A = 0.25$, and the same for ENM with $w_A = 0.5$. Also, shows the same behaviour for ENM with $w_A = 0.0$ vs ENM with $w_A = 0.25$, and the same vs ENM with $w_A = 0.25$. Interestingly, there is no significant difference between ENM with $w_A = 0.25$, and ENM with $w_A = 0.5$.

vs $w_A = 0.5$.

These results show the impact of the aesthetic criteria by finding a better article’s placement, in some cases similar to the In-place magnification if it is aesthetically pleasing. Also, the aesthetic term can modify some shapes to find better layouts, without impacting the quality of the headlines. Regarding the aesthetic criteria, in most cases, our method generates a layout without any line excess, and in other examples, the set of original shapes results is better than any other set of alternative shapes, leading to a redistribution of the original shapes.

5.5. Qualitative Analysis of the Generated Pages

Finally, we propose a qualitative analysis of the generated pages.

Our first analysis consists of observing the best pages found associated with the 20 inputs from our dataset. Our observations are summarized in

Table 7. Results show different behaviors depending on the input characteristics (number of alternative shapes, size of big articles, number of articles, or geometrical issues difficult to quantify).

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
Different shapes than input?		x		x	x	x	x		x	x	x	x	x	x	x					
Different positioning than input?	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Difference of shapes between $w_A = 0$ and $w_A = 0.25/0.5$?		N		S	M	B	M		N	N	S	N	N	N	M					
If no difference in shapes, is the article's placement when $w_A = 0$ similar to input?																x				
If no difference in shapes, is the article's placement when $w_A = 0.25/0.5$ similar to input?	x		x					x								x	x	x		x

Table 7: Characteristics of the rendered layouts for each input. N states for *None*, S for *Small*, M for *Moderate* and B for *Big*. We highlight in grey the layouts for which we show the results in Fig. 11–14 as they are representative of the different behaviors we get.

Figures 11–14 show different types of behaviors, leading to three conclusions: (1) Using aesthetic criteria can preserve the quality of the in-place magnification or gives results that improve this term, (2) different values of w_A do not necessarily give better aesthetic improvement since there are geometrical constraints that limit this term and (3) same values of $H(\cdot)$ do not mean equivalent layouts, since we saw that this term doesn't change between different values of w_A , but produces a different set of shapes in some cases. This might lead to future work about quantifying differences between layouts in a more detailed way.



(a) Input (b) In-place mag. (c) $w_A = 0$ (d) $w_A = 0.25$ (e) $w_A = 0.5$

Figure 11: Results for p_1 . The method did not find a better set of alternative shapes because the original set is already near-optimal. Also, when we introduce aesthetic criteria, the results turn similar to the in-place magnification which turns better in aesthetic terms than (c) which is a completely different distribution of articles.



(a) Input (b) In-place mag. (c) $w_A = 0$ (d) $w_A = 0.25$ (e) $w_A = 0.5$

Figure 12: Results for p_2 . The method found a better set of shapes in terms of $H(\cdot)$, and it is the same across different values of w_A . Again, introducing aesthetic criteria changes the distribution, with no differences between (d) and (e).

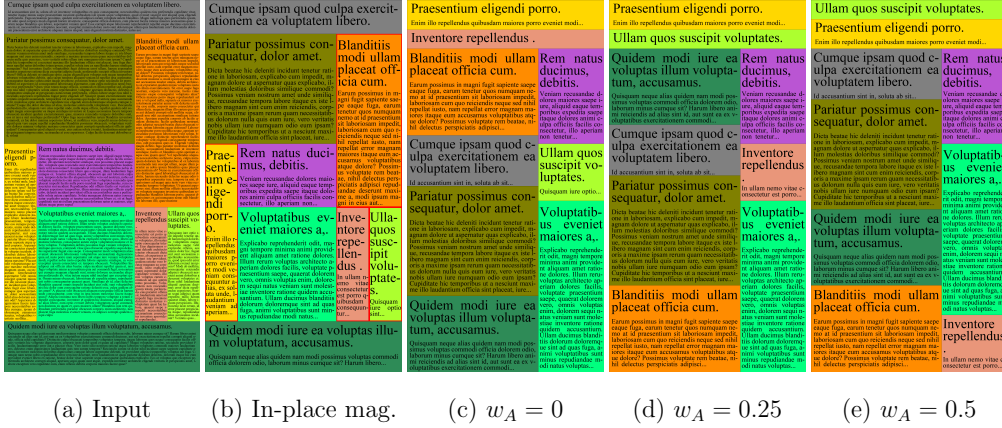


Figure 13: Results for p_4 . Here we see that different values of w_A change the set of shapes, without losing quality of $H(\cdot)$, and increasing aesthetic quality.

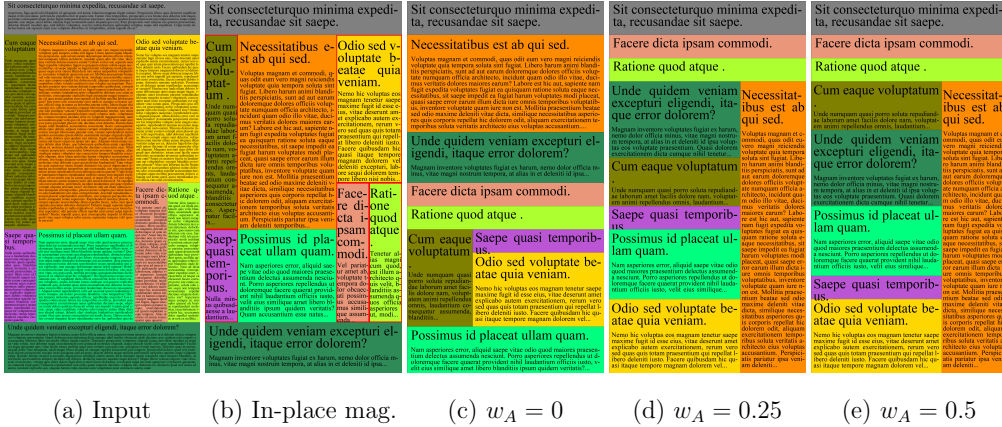


Figure 14: Results for p_6 . Similarly to p_4 , we see that different values of w_A change the set of shapes, without losing quality of $H(\cdot)$, and increasing aesthetic quality. In this case, the change of shapes is bigger than for p_4 .

6. Conclusion

Seeking a solution to magnify newspaper pages while preserving headline quality (readability and usability) led us to define a novel and nontrivial layout generation and packing problem called Newspaper magnification with preserved entry points. The major difference of this problem w.r.t. other similar packing problems is that articles can have any width (and height), as soon as their size is preserved.

To define our objective as optimization functions for this novel problem, we defined an energy that considers two aspects related to the number of lines in headlines and aesthetic criteria, applying some weight to the latter one considering that our main focus is the quality of headlines. This allows for improving headlines while introducing additional fitness criteria that define the quality of the positioning of each article.

In this paper, we proposed to solve this problem thanks to an evolutionary algorithm based on packing techniques, allowing us to generate magnified layouts in a fast way and with a simple encoding/decoding process. In the first part of our experiments, we checked the variability of execution time and energy related to different seeding for each random procedure in the algorithm, together with the selection of three different heuristics to pack articles. Results show that there is no significant variability between seeds neither in execution time and $E(.)$ value, which proves the algorithm's stability and allows running our method only once per input. Regarding differences between heuristics, we found significant differences only in execution time, where BLF arises as the fastest, followed by BF and finally iBL. In our second part of the experiments, we showed the difference between in-place magnification

and our best layouts for difference values of aesthetic term weight w_A . We demonstrated the impact of the aesthetic criteria by finding a better article’s placement, in some cases similar to the input if it is aesthetically pleasing. Also, interestingly, our method is capable of modifying some shapes to find better layouts without impacting the quality of the headlines.

Overall, we believe the Newspaper magnification with preserved entry points task will motivate the community along the lines of research for document engineering, user interface, and accessibility. Indeed, it has potential application for several use cases which are related to reading comfort on tablets. A first use case that motivated our study concerns low-vision people for whom newspapers pose a unique challenge due to their unpredictable layout and condensed formatting [13], making them hardly accessible. A second use-case concerns in fact all readers willing to read their newspaper on a tablet more comfortably without being bothered by the manipulations associated with the pinch-and-zoom actions. Of course, these applications call for more research to target real newspaper situations, such as (1) considering other entry points than headlines (e.g., images), (2) defining enhanced computational aesthetics criteria that also consider the content and not only the organization of the articles, (3) focusing more on the algorithmic performance to propose fast magnification, and (4) behavioral studies to assess the interest of the new magnification principle depending on the use-case.

Acknowledgements

We are thankful to B. Génuit (Demain Un Autre Jour, Toulouse, France) who provided expertise that assisted the research, and Dr. Hui-Yin Wu

(Université Côte d’Azur, Inria, France) who shared with us her annotation tool that we used to build our dataset.

References

- [1] Alvarez-Valdes, R., Parreño, F., and Tamarit, J. (2008). Reactive grasp for the strip packing problem. *Computers and Operations Research*, 35(4):1065–1083.
- [2] Baker, B., Coffman, E., and Rivest, R. (1980). Orthogonal packings in two dimensions. *SIAM Journal of Computing*, 9:846–855.
- [3] Brabrand, G. B. (2008). Dynamic layout optimization for newspaper web sites using a controlled annealed genetic algorithm.
- [4] Burke, E. K., Hyde, M., Kendall, G., and Woodward, J. (2010). A genetic programming hyper-heuristic approach for evolving 2-d strip packing heuristics. *IEEE Transactions on Evolutionary Computation*, 14(6):942–958.
- [5] Burke, E. K., Kendall, G., and Whitwell, G. (2004). A new placement heuristic for the orthogonal stock-cutting problem. *Operations Research*, 52(4):655–671.
- [6] Chazelle, B. (1983). The bottom-left bin-packing heuristic: An efficient implementation. *IEEE Transactions on Computers*, C-32(8):697–707.
- [7] Errea, J. and Gestalten, editors (2018). *Newspaper Design: Editorial Design from the World’s Best Newsrooms*. gestalten.

- [8] Fortunati, L., Taipale, S., and Farinosi, M. (2015). Print and online newspapers as material artefacts. *Journalism*, 16(6):830–846.
- [9] Garcia, M. R. and Stark, P. (1991). *Eyes on the News*. Poynter Inst Media Studies.
- [10] Gautier, D. and Gautier, C. (2017). *Design, Typography etc.: A Handbook*. niggli.
- [11] Geigel, J. and Loui, A. C. P. (2000). Automatic page layout using genetic algorithms for electronic albuming. *Internet Imaging II*, 4311(August 2002):79–90.
- [12] Harrington, S. J., Naveda, J. F., Jones, R. P., Roetling, P., and Thakkar, N. (2004). Aesthetic measures for automated document layout. In *Proceedings of the 2004 ACM Symposium on Document Engineering, DocEng '04*, page 109–111, New York, NY, USA. Association for Computing Machinery.
- [13] Hersh, M. and Johnson, M. (2010). *Assistive Technology for Visually Impaired and Blind People*. Springer, London.
- [14] Hofmeister, J., Heller, D., and Radach, R. (1999). *The Return Sweep in Reading*, pages 349–357. Springer US, Boston, MA.
- [15] Hollander, B. A., Krugman, D. M., Reichert, T., and Avant, J. A. (2011). The e-reader as replacement for the print newspaper. *Pub Res Q*, 27:126–134.

- [16] Holsanova, J., Rahm, H., and Holmqvist, K. (2006). Entry points and reading paths on newspaper spreads: comparing a semiotic analysis with eye-tracking measurements. *Visual Communication*, 5(1):65–93.
- [17] Ihlström, C., Åkesson, M., and Nordqvist, S. (2004). From print to web to e-paper—the challenge of designing the e-newspaper. *Proceedings of ICC 8th International Conference on electronic Publishing*, page 249–260.
- [18] Legge, G. E. (2007). *Psychophysics of reading in normal and low vision*. Psychophysics of reading in normal and low vision. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US. Pages: xvi, 229.
- [19] Liu, D. and Teng, H. (1999). An improved bl-algorithm for genetic algorithm of the orthogonal packing of rectangles. *European Journal of Operational Research*, 112(2):412–420.
- [20] Ngo, D. C. L., Teo, L. S., and Byrne, J. G. (2000). A mathematical theory of interface aesthetics. *Visual Mathematics*, 10(8).
- [21] Ngo, D. C. L., Teo, L. S., and Byrne, J. G. (2003). Modelling interface aesthetics. *Information Sciences*, 152(SUPPL):25–46.
- [22] Oliveira, J., Neuenfeldt Júnior, A., Silva, E., and Carravilla, M. (2016). A survey on heuristics for the two-dimensional rectangular strip packing problem. *Pesquisa Operacional*, 36:197–226.
- [23] Purvis, L. (2002). A genetic algorithm approach to automated custom document assembly. pages 131–136.

- [24] Thomas, J. and Chaudhari, N. (2014). A new metaheuristic genetic-based placement algorithm for 2d strip packing. *Journal of Industrial Engineering International*, 10.
- [25] Vasilev, M. R., Adedeji, V. I., Laursen, C., Budka, M., and Slattery, T. J. (2021). Do readers use character information when programming return-sweep saccades? *Vision Research*, 183:30–40.

7. Appendix

7.1. Line Splitting Criteria

Considering that we have a fixed bounding box given by the shape, the spaces between the words on each line are adjusted to achieve left-right justification such that the sentence snugly fits into the sentence bounding box ([18]). Here, we impose as a rule that each space may be narrowed to no less than 80% or widened by no more than 125% of the width of a space. By this, we avoid excessively loose or tight spacing between the words on each line to maintain the legibility of the sentence. Considering this, we define the line-splitting criteria: if the words are too close among others (i.e., each space in a single line is narrowed to less than 80% of the normal size), the sentence is broken in two (or more, to fulfill this constraint). The width of each line of text is, of course, given by the width of the bounding box. Note that any other line-splitting rule could be used without losing the generality of the proposed method.

7.2. Method to Measure Alignment

We use the approach described in [12]. Let us focus on the horizontal alignment of the left edges to explain the method, that can be applied analogously to the top. First, we sort in increasing order all the left edge x -coordinates of each shape to obtain a list $\{x_1, \dots, x_N\}$. For two consecutive x -coordinates (x_i, x_{i+1}) , we define the cost:

$$g(x_i, x_{i+1}) = \frac{Q_1}{Q_1 + |x_{i+1} - x_i|}, \quad (3)$$

where Q_1 is a parameter. If the left edges of \mathcal{L}_i and \mathcal{L}_{i+1} are aligned, then $g(i, i+1) = 1$. Considering all consecutive edges, we can define the horizontal

alignment measure for left edges by:

$$A_l(\mathcal{L}) = \frac{\sum_{i=1}^{N-1} g(x_i, x_{i+1})}{N-1},$$

Analogously, we can define an alignment measure top edges ($A_t(\mathcal{L})$). Then, we define $A_1(\mathcal{L})$ as:

$$A_1(\mathcal{L}) = \frac{A_l + A_t}{2} \in [0, 1].$$

7.3. Method to Measure Regularity

We use the approach described in [12]. Let us focus on the horizontal alignment of left vertical edges to explain the method; first, we obtain and sort all alignment points $\{x_1, \dots, x_N\}$ in the same way explained for $A_1(\cdot)$. Now we are interested in changes in the alignment points, not in the points themselves, and also we are not interested when $x_i = x_{i+1}$ because we are trying to see the pattern of changes between different alignment points. In this case, we measure the distance $z_{i,i+1}$ between two consecutive and different points, and save these values for the next step. In Fig. 15 we explain this approach using a histogram to illustrate what are we measuring.

Having the distance $z_{i,i+1}$, we follow the same strategy explained in the definition of $A_1(\cdot)$, but using this $z_{i,i+1}$ values instead of the left edge x-coordinates. That means (1) sort in an increasing order all the $z_{i,i+1}$ values, and (2) for two consecutive $z_{i,i+1}$ values ($z_{i,i+1}, z_{i+1,i+2}$), we define the cost:

$$f(z_{i,i+1}, z_{i+1,i+2}) = \frac{Q_2}{Q_2 + |z_{i+1,i+2} - z_{i,i+1}|}, \quad (4)$$

where Q_2 is a parameter. The maximal value is 1 when $z_{i,i+1}$ and $z_{i+1,i+2}$ have the same value. Considering all $z_{i,i+1}$, we define the horizontal regularity

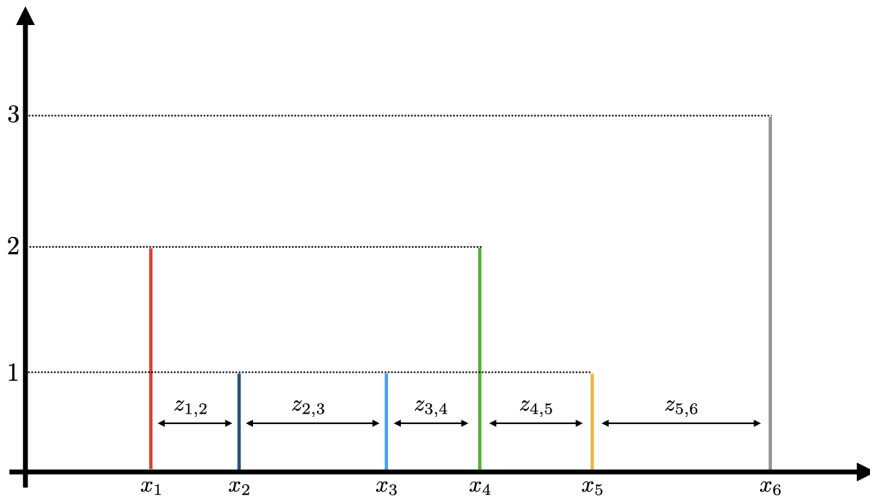


Figure 15: Illustration of the regularity measure, using a histogram for visualization: Each entry in the histogram shows the value of an alignment point. Then, the distance between two consecutive bars corresponds to $z_{i,i+1}$.

measure by:

$$R_h(\mathcal{L}) = \frac{\sum_{i=1}^{Z-1} f(z_{i,i+1}, z_{i+1,i+2})}{Z-1},$$

where Z is the number of $z_{i,i+1}$ values.

Analogously, with horizontal regularity (computed using left-edge alignment points) R_h and vertical regularity (computed starting from top-edge alignment points) R_v , compute the overall regularity measure:

$$A_2(\mathcal{L}) = \frac{R_h + R_v}{2} \in [0, 1].$$

7.4. Method to Measure Balance

Here we use the approach described in [21]. Let us consider first the balance between the left and right sides of the page. Given the vertical line \lceil splitting the page into two equal parts (\mathcal{D}_l and \mathcal{D}_r), we first define the left weight by doing a weighted sum of the areas of all $\{\mathcal{L}_i\}_{i=1,\dots,N}$ whose centers calculated as $(x_i + \frac{w_i}{2}, y_i + \frac{h_i}{2})$ belong to \mathcal{D}_l , i.e.,

$$w_l(\mathcal{L}) = \sum_{\mathcal{L}_i / (x_i + \frac{w_i}{2}, y_i + \frac{h_i}{2}) \in \mathcal{D}_l} a(\mathcal{L}_i) \delta_i,$$

where δ_i denotes the distance between the center of each article and line \lceil , and $a(\mathcal{A}_{i,j})$ corresponds to article's area. Doing the same for the right-hand side, we can estimate a weight w_r and then define the balance between the left and right sides of the page as:

$$B_{lr}(\mathcal{L}) = \frac{|w_l - w_r|}{\max(|w_l|, |w_r|)},$$

which is a scalar value in $[0, \dots, 1]$. The same procedure can be applied to estimate the balance between the top and bottom sides of the page (B_{tb}), leading to our final balance term:

$$A_3(\mathcal{L}) = 1 - \frac{B_{lr} + B_{tb}}{2} \in [0, 1].$$