Content Extraction from Marketing Flyers

Ignazio Gallo, Alessandro Zamberletti and Lucia Noce

University of Insubria,
Department of Theoretical and Applied Science,
Via Mazzini, 5, 21100 Varese, Italy
{ignazio.gallo,a.zamberletti,lucia.noce}@uninsubria.it
http://artelab.dicom.uninsubria.it/

Abstract. The rise of online shopping has hurt physical retailers, which struggle to persuade customers to buy products in physical stores rather than online. Marketing flyers are a great mean to increase the visibility of physical retailers, but the unstructured offers appearing in those documents cannot be easily compared with similar online deals, making it hard for a customer to understand whether it is more convenient to order a product online or to buy it from the physical shop. In this work we tackle this problem, introducing a content extraction algorithm that automatically extracts structured data from flyers. Unlike competing approaches that mainly focus on textual content or simply analyze font type, color and text positioning, we propose novel and more advanced visual features that capture the properties of graphic elements typically used in marketing materials to attract the attention of readers towards specific deals, obtaining excellent results and a high language and genre independence.

Keywords: Content Extraction, Portable Document Format, Visual Features, Marketing Flyers.

1 Introduction

Although e-commerce has been increasing in popularity in recent years, unstructured documents such as marketing flyers and advertising emails are still effective ways to promote products and special offers. In this study we propose a novel content extraction algorithm to automatically extract entities of interest from marketing flyers containing commercial product offers.

Most of the deals appearing within commercial flyers refer to physical retailers, and the data that can be gathered from those marketing documents is particularly appealing to online price comparison shopping engines to fill the existing gap between online and physical shopping. In fact, a searchable collection containing both physical deals extracted from marketing flyers and offers gathered from online listings would allow customers to determine whether it is more convenient to order a product online and wait for order preparation, shipping and delivery, or to physically drive to the retailer location and buy it

straight from the shelf. This represents a great opportunity for physical retailers to compete against online marketplaces.

The PDF standard is increasingly being used by physical retailers to create marketing materials that maintain the same visual characteristics on every different device. This is particularly important, as commercial flyers typically contain many graphic elements specifically designed to let customers quickly understand the positions of relevant offers within the pages or to attract their attention towards particular deals or sections. Most content extraction works in literature extract entities of interest from structured or semi-structured documents relying exclusively on textual features, ignoring the visual characteristics of the processed documents [1–5]. While these approaches may obtain satisfying results when processing heavily structured or plain text documents, recent works have shown that, when dealing with less structured documents containing lots of graphic elements, textual features are not discriminative enough to accurately identify all the entities of interest [6, 7]. In fact, visual characteristics are typically used to highlight and categorize paragraphs of text, and therefore represent a large amount of information that can and should be used to more accurately distinguish entities of interest having low discriminative textual contents.

Casting the problem to PDF documents, existing studies transform the processed PDF content streams into raw text, wasting the visual formatting information contained within the documents, such as font type, size, color and text positioning [4, 5]. If we delve even deeper into the world of PDF documents and we focus on PDF marketing flyers, it has been proved that combining textual information with additional features describing the visual characteristics of the deals in the processed documents increases classification performances [6].

Similarly to [6], the goal of this study is to underline the important role that visual features have in distinguishing different entities of interest within highly unstructured but visually rich documents. Unlike other works in literature, we propose novel and more complex visual features, *e.g.* font type frequency, markup density, xObject presence, *etc.*, that capture a wider set of visual characteristics of the processed documents and extract them directly from the PDF content streams to avoid visual distortions that may arise when converting PDF to other formats such as HTML. The use of a limited set of textual features, combined with highly discriminative visual features, enables our system to obtain satisfying results on visually dissimilar marketing flyers, while also maintaining a high language and genre independence. Although in this study we focus exclusively on marketing materials, the proposed visual features are not handcrafted for that specific domain; as such, they can be used in a wide variety of contexts in which the visual characteristics of the processed documents are discriminative for the entities that need to be identified.

2 Related Works

In the following paragraphs we introduce some recent works on content extraction that combine textual features with visual features describing the visual

(a) Original (b) Converted

Fig. 1: The conversion of a PDF file (a) to an HTML format (b) may generate formatting alterations that substantially change the visual characteristics of the original document. Conversion tool: PDF2HTML [6].

elements that typically help humans to quickly understand the different entities within the analyzed documents.

Radek *et al.* [8, 9] propose a HTML content extraction method based on a page segmentation algorithm that splits rendered HTML pages into multiple basic areas that are visually separated from each other due to different background colors, frames or markup separators. Areas having similar visual characteristics are then clustered together into semantically correlated blocks which are assigned to different classes of interest on the basis of their font, spatial, text and color features.

Sun *et al.* [10] use Content Extraction via Text Density (CETD) to detect entities of interest within web pages, based on the observation that content text typically contains long and simply formatted sentences, while noise (navigation panels, advertisements, copyright information, disclaimer notices, *etc.*) is highly formatted, contain less text and shorter sentences. Statistical information about hyperlinks within the HTML Document Object Model (DOM) tree are also taken into account during the classification process. Even though CETD is completely language independent, it achieves and outperforms several competing content extraction algorithms, showing that an analysis of the lexicon used in the processed documents is not mandatory to obtain state-of-the-art results.

The idea that is possible to identify entities of interest in a language independent way is further inspected in [7]. In this work, the authors propose an automated system for content extraction from HTML web pages; the algorithm requires no user interaction and it relies exclusively on visual features extracted from both HTML tags and CSS style sheets.

Apostolova and Tomuro [6] use visual features like font size, font color and text y coordinate to detect entities of interest from online PDF flyers of real estate offerings, obtaining significant improvements over the detection results

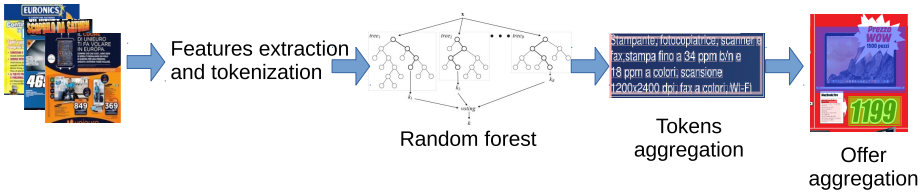

Fig. 2: Pipeline of the proposed method.

achieved using standard textual features. To extract visual attributes, PDF flyers are firstly automatically converted to HTML by analyzing PDF format operators in the PDF content stream; style attributes are then dynamically extracted from the document by rendering the page with Google Chrome browser. [1]

The quality of the HTML document automatically generated when converting a PDF document depends on the degree of accordance between the PDF that is being converted and the ISO 32000-1/32000 specifications published by Adobe, *e.g.* as shown in Figure 1, when *BeginText* (BT), *EndText* (ET) and other format operators are misused in the PDF content stream, the HTML page resulting from the automatic conversion is flawed; the same holds when the processed PDF has been generated using a distiller which does not comply to the published ISO 32000-1/32000 specifications. [2]

Although none of the aforementioned algorithms extract content directly from PDF documents, it is possible to compute visual features from PDF content streams without having to convert them into HTML documents, thus avoiding the risk to alter the original document visual formatting style. In this study, multiple visual feature planes conforming with the formatting style of the original processed PDF document are automatically built by processing the format operators found in the content stream. Figure 3 shows how those visual features maintain the correct visual formatting style even for a PDF marketing flyer that cannot be automatically and correctly converted to HTML.

3 Proposed Method

The processing pipeline of the proposed approach is presented in this section: (i) feature planes and textual information are extracted from the processed marketing flyer; (ii) visual information gathered from feature planes are used to classify each word within the processed flyers using a properly trained random forest classifier; (iii) neighbouring words having similar visual characteristics are merged into semantically correlated paragraphs; (iv) paragraphs representing correlated product titles, descriptions and prices are further merged together to identify the deals contained within the processed flyer. The whole pipeline is summarized in Fig. 2 and described in detail in the remainder of this section.

[1] http://pdftohtml.sourceforge.net/
[2] http://www.adobe.com/devnet/pdf/pdf_reference.html

(a) original (b) black text plane (c) color plane (d) markup plane

Fig. 3: Feature planes (b,c,d) automatically computed for a given flyer (a).

3.1 Visual Features

Given a PDF document, we compute the following visual feature planes: *black text*, *colored text* and *markup* (examples are provided in Fig. 3).

The *black text* plane isolates all the selectable text elements, and it is obtained by removing all but font, position and begin/end text PDF format operators from the original PDF content stream. As shown in Fig. 3b, all the selectable text elements are black and placed on a plain white background. Text position and formatting are preserved as in the starting document.

The *black text* plane is used to recognize text characters that are not readable due to missing font character maps (CMAP). In fact, when a font CMAP is missing, it is not possible to obtain the unicode representation of the fonts glyphs unless they are visually recognized using an appropriate OCR engine [11].

The *colored text* plane is obtained by removing all the following PDF format operators from the PDF content stream: b, B, b*, B*, BDC, BI, BMC, BX, c, d, Do, DP, EI, EMC, EX, f, F, f*, G, g, h, i, ID, j, J, l, m, M, MP, n, re, ri, s, S, sh, v, w, W, W*, y. Detailed information of the meaning of each PDF format operator are available in the ISO 32000-1/32000 specification document. As shown in Fig. 3c, the resulting plane is a copy of the original document, in which all the text components have been isolated from other PDF elements such as images, markup and background patterns. Unlike the *black text* plane, in the *colored text* plane text components maintain their original color, shadow, opacity and line style properties.

The *markup* plane contains exclusively all the markup and geometric elements used in the original document. As shown in Fig. 3d, this includes both text overlays (highlight, strikethrough, underline, *etc.*) and geometric figures drawn using PDF format operators, such as re and v (rectangles and lines respectively). In visually rich marketing documents, these geometric elements are particularly important as they are typically used to isolate semantically correlated text elements and draw the attention of customers towards relevant deals.

Table 1: List of features computed for each token in the processed flyer.

Feature	Description
is_number	Boolean value representing whether the token contains only digits.
digits_percentage	The percentage of digits in the token.
all_upper_case	Boolean value representing whether the token contains only upper case letters.
only_first_upper_case	Boolean value indicating whether only the first letter of the token is capitalized.
token_font_size	The largest font size of text characters in the token, measured in pixels.
token_angle	The average orientation angle of text characters in the token, measured in degrees.
token_position	The normalized position (x, y) relative to the page size.
token_color	The 3 most recurring RGB values in the region occupied by the token within the *colored text* plane.
token_font_frequency	The frequency of a specified font F_i in a specified page.
token_color_frequency	The frequency of token_color within the current document page.
token_markup_color_frequency	The frequency of the most recurring RGB value in the region occupied by the token within the *markup* plane.
token_font_page_frequency	The frequency of the token font measured over all the pages in the current document.

The previously described visual feature planes are used to compute the salient visual aspects of each token, or word, within the processed document. In many languages using Latin alphabet, word space may be used as a good approximation of a word divider. Unfortunately, using the space character as a delimiter to extract words from a PDF file does not always lead to optimal results. In fact, unlike other types of structured documents, PDF files do not always store word spaces, *e.g.* in some PDF documents, multiple BT/ET blocks (*begin text* and *end text* respectively) may be used within the same paragraph of text with different transformation matrix; in these cases, even though the appearance of the resulting PDF files is correct, some spaces between words are missing in the plain text stream extracted from the PDF content stream. To overcome this issue, in our method we compute the average width of text characters within the processed PDF, and use that as a metric to split text into tokens/words.

For each token we compute the set of visual features described in Table 1. Simple features such as font size, orientation, *etc.*, can be computed directly for each token, without having to analyze the entire document. On the other hand,

Table 2: Relevant entities for the extraction of offers from marketing flyers.

Entity	Description
Title	The insertion/offer title. It is usually composed of the brand name, the product name and some product specifications.
Description	The description of the object specified in the title. It gives additional details about the item.
Price	The final item price. All the strikethrough prices that may be associated with the same item need to be ignored, only the definitive price tag has to be taken into account.
Other	Content that is not related to offers or deals. This may include, but is not limited to, the flyer's title, expiration date, disclaimers and product tags.

the computation of more advanced features such as token font, color and markup frequencies require to analyze the entire content of the processed document.

More in detail, given a token/word t with its font f_t in a page p, the Token Font Frequency TFF of t in p is computed as follows:

$$TFF_{t,p} = \frac{n_{f_t}}{|p|} \tag{1}$$

where n_{f_t} is the number of tokens having font f_t in p, and $|p|$ is the total number of tokens in p.

Similarly, the Token Color Frequency TCF of a token t having font color c_t in p is computed as follows:

$$TCF_{t,p} = \frac{n_{c_t}}{|p|} \tag{2}$$

where n_{c_t} is the number of tokens having font color c_t in the page p.

Due to shadows, opacities and different text characters colors, a token/word may have multiple colors within the processed document. In our pipeline, for each token, we exploit the *colored text* plane to identify the most recurring RGB color appearing within the visual region that the token occupies in the page p, and use that RGB value as c_t.

A smilar process is used to compute the Token Markup Color Frequency $TMCF$ of the token t within the page p:

$$TMCF_{t,p} = \frac{n_{m_t}}{|p|} \tag{3}$$

where n_{m_t} is the number of tokens having markup color m_t in p. The value of m_t is computed as the most recurring RGB color in the region occupied by t in

the *markup* plane for p. Please note that during the computation of token color c_t and markup color m_t, the white background on which all the elements are placed on the 3 feature planes is ignored.

Marketing flyers may have multiple pages, in these cases it is interesting to analyze the Font Page Frequency FPF of a token t in the whole document d:

$$FPF_{t,d} = \frac{|\{p_i : f_t \in p_i\}|}{|d|} \tag{4}$$

where $|\{p_i : f_t \in p_i\}|$ is the number of pages containing the font f_t in the marketing flyer d, and $|d|$ is the total number of pages in d.

Font Page Frequency is particularly relevant when processing marketing documents having multiple pages, as the fonts used to denote entities of interest typically do not change from page to page. This means that fonts having a high Font Page Frequency value are typically associated with interesting content, while low frequency fonts are usually associated with noise content, such as footnotes or disclaimers.

3.2 Token Classification and Aggregation

Tokens extracted from the processed flyers are classified as belonging to one of the entities of interest listed in Table 2. The classification task is carried out using a Random Forest classifier from Waikato Environment for Knowledge Analysis (WEKA) [12] library. Random Forest classifiers perform as well as SVM or NN classifiers when trained using a sufficient amount of data, while also being significantly faster to train [13].

Once every token in the page has been assigned to a class of interest, they need to be aggregated to form products titles, descriptions and prices. This aggregation task is carried out using an ad-hoc clustering algorithm that takes into account both the class and the position of tokens within the processed flyer.

At its first iteration, the algorithm selects the bounding box of a random seed token classified as belonging to either Title, Price or Description class, and tries to join that bounding box with all the other neighbouring bounding boxes of tokens classified as belonging to the same class c that are located at a distance $d < \epsilon$. This newly formed bounding box is then added to the page in place of all the joined bounding boxes. At each iteration a new seed token, that has not been previously selected, is chosen. The algorithm stops when all the tokens have been aggregated.

The result of this merging phase is a set of bounding boxes that represent titles, descriptions and prices of all the offers available on the flyer (see Fig. 4).

3.3 Offer Aggregation

The offers extraction process from each flyer is the last step of the presented method. This is not a trivial task as it cannot be carried out simply by considering the minimum distance between the various elements that form an offer. In

Fig. 4: Examples of flyers manually tagged by experts. The relevant entities listed in Table 2 are highlighted as coloured rectangles (Title, Description and Price).

fact, there are many cases in which one or more of the bounding boxes for the 3 relevant elements that make up an offer (Price, Title and Description) are visually closer to the bounding boxes of elements from another offer. In such cases, clustering exclusively on the basis of the distance between different bounding boxes does not lead to optimal results.

A better approach consists in clustering the bounding boxes in such way that the coverage provided by the final clusters over the processed page is maximized. This approach is motivated by the fact that marketing documents do not usually have many void areas, because retailers typically try to lower printing costs by adding as many offers as they can within each page to reduce the total size of the final flyer. As such, each offer within a page is usually localized in a particular area, and its bounding box has a minimal overlap with the other offers.

The textual information associated with an offer O is a triple (T, D, P) composed of a Title T, a Description D and a Price P. In this work, product images are not taken into account because finding the correct association between an image and its respective textual description requires a specific study, which is out of the focus of this work.

As previously stated, our offer aggregation algorithm tries to minimize the intersection area between all the bounding boxes for the offers in the processed page. The algorithm starts by selecting the bounding box of a random Price P_i and merges it with its closest Description D_i and Title T_i bounding boxes to form an offer hypothesis O_i. The same process is repeated for all the remaining Prices in the page to form a finite set of hypotheses $H_{P_i} = \{O_0, \ldots, O_n\}$. The sum $S_{H_{P_i}}$ of the intersection areas between the bounding boxes for the offers in H_{P_i} is then calculated as follows:

$$S_{H_{P_i}} = \sum_{j \neq k} O_j \cap O_k, \;\; j, k \in \{0, \ldots, n\} \tag{5}$$

This whole process is repeated multiple times, each time changing the starting seed Price, until all the Prices in the flyer have been selected as initial seeds. The set of offer hypotheses having minimum intra-intersection area is then selected as the best one.

4 Experiments

In the remainder of this section we present the experimental results obtained testing the proposed method on marketing flyers randomly collected from different retailers. Throughout our experimental activity we evaluate quantitatively the accuracy of the method both at identifying and classifying entities of interest within the processed flyers; and at aggregating the detected entities into offers.

4.1 Dataset

In order to evaluate the proposed approach, a total number of 1194 product offers have been gathered from 197 marketing flyers produced by 12 different retailers. The collected documents come from heterogeneous domains (electronics, gardening, clothing, *etc.*) and present substantially different design styles.

Each flyer has been manually labelled by a team of 4 experts using a specially designed GUI. As shown in Fig. 4, the experts were instructed to provide both the coordinates of all the product Titles, Descriptions and Prices in the pages, and the associations between those bounding boxes and the different offers within the pages. The information gathered from the different experts has been averaged to obtain the final ground-truth data used to evaluate the proposed method.

4.2 Evaluation Metrics

We evaluate the accuracy of the method both at classifying/aggregating individual tokens, and at aggregating the merged tokens into product offers.

Since our ground-truth data is composed of labelled bounding boxes manually drawn by experts over the different flyers, we measure the accuracy of the proposed approach by evaluating the intersection-over-union (IoU) [14] score between the bounding boxes detected by the proposed approach and the respective ground-truth information.

Each entity is evaluated independently from the others. Given a page with its ground-truth data for one of the entities from Table 2, and the aggregated predictions provided by the model for the same entity class; the evaluation process for the token classification/aggregation phase is carried out as follows: we compare the IoU score between each ground-truth bounding box and the predictions provided by the model; if one of the predicted bounding boxes achieves an IoU score greater than 0.5 with the ground-truth bounding box, the prediction is considered correct. For every ground-truth bounding box at most one predicted bounding box might be considered correct. Given the number of correct predictions, we compute the classic Precision, Recall and F-measure values.

Given a page with its ground-truth offer data, and the offer hypotheses generated as in Sec. 3.3, the evaluation process for the offer aggregation phase is carried out as follows: we compare the IoU score between each component of a ground-truth offer (Title, Description and Price) and the bounding boxes for the same component in the offer hypotheses; if every predicted component for a given offer hypothesis has an IoU that is greater than 0.5 with its respective ground-truth offer component, then the predicted offer is considered correct. For every ground-truth offer at most one hypothesis might be considered correct.

4.3 Results

Starting from a total number of 51045 tokens extracted from the dataset, 70% are used for training the classifier and the remaining 30% for testing. For each token, we compute the features listed in Table 1 using a 3×1 contextual sliding window centered on the token. As such, each token is classified on the basis of both its attributes and the attributes of its left and right neighbours.

The first step in the evaluation process aims at detecting the importance of the features listed in Table 1 using Information Gain [12]. The first 20 features selected by Information Gain in order of importance, are: font frequency, font size, font frequency right, font frequency left, font size left, font size right, y position, y position left, 1st RGB color left, 1st RGB color right, font page frequency, font page frequency right, font page frequency left, 2nd RGB color, 3rd RGB color, digit percentage, all upper case, 2nd RGB color left. Using Information Gain, only 2 among the top 20 ranked features are textual, underlining the importance of visual features.

The second experiment aims at measuring the goodness of the proposed Random Forest classifier. As listed in Table 2, tokens may belong to one of four possible classes: Title, Description, Price and Other In our experiment the classifier may contain a maximum of 10 trees; no limitations are posed on the depth of each tree; each tree considers 6 random features. The confusion matrix is presented in Table 3, 93.36% of the patterns are correctly classified, with a k-value of 0.89.

With the last experiment we evaluate the phases described in Sec. 3.2 and 3.3: the aggregation of tokens, and the subsequent aggregation of merged tokens into product offers. We measure Precision, Recall and F-measure values achieved on test set, while varying the token aggregation threshold ϵ from $0.1 \cdot \text{token\_height}$ to $10 \cdot \text{token\_height}$. We report the best obtained results in Table 3; they have been obtained setting $\epsilon = (2 \cdot \text{token\_height})$. The accuracy of the offer aggregation phase is influenced by the errors committed during the previous token classification/aggregation phase. Since the proposed method is structured as a waterfall of steps, this accuracy decreasement is inevitable.

5 Conclusion

An ad-hoc method for the automatic extraction of structured product offers from marketing flyers has been proposed. The presented approach heavily re-

Table 3: Evaluation of the proposed method. On the left, the confusion matrix for the Random Forest classifier (k-value: 0.89). On the right, the results of both the token classification/aggregation (Descr., Title and Price) phase, and the offer aggregation phase (Aggr. offers). The best results were obtained by setting the token aggregation threshold to $\epsilon = 2 \cdot \text{token\_height}$.

	Descr.	Title	Price	Other
Descr.	95.39%	5.57%	3.70%	3.27%
Title	2.70%	91.51%	2.94%	3.51%
Price	0.19%	0.67%	87.31%	2.04%
Other	1.72%	2.25%	6.05%	91.18%

	Precision	Recall	F-measure
Description	0.740	0.655	0.695
Title	0.789	0.837	0.812
Price	0.815	0.916	0.862
Aggr. offers	0.487	0.547	0.515

lies on novel visual features that capture the formatting details typically used in marketing documents to accurately distinguish relevant entities within the processed flyer's pages. The method has been evaluated over a collection of randomly collected flyers, achieving satisfying results while also maintaining an excellent language and genre independence due to the limited use of classical textual features.

References

1. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Linguisticae Investigationes **30** (2007) 3–26
2. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: CoNNL. (2009) 147–155
3. Ling, X., Weld, D.: Fine-grained entity recognition. In: AAAI. (2012)
4. Yuan, F., Liu, B., Yu, G.: A study on information extraction from pdf files. In: ICMLC. (2006) 258–267
5. Prokofyev, R., Demartini, G., Cudré-Mauroux, P.: Effective named entity recognition for idiosyncratic web collections. In: WWW. (2014) 397–408
6. Apostolova, E., Tomuro, N.: Combining visual and textual features for information extraction from online flyers. In: EMNLP. (2014) 1924–1929
7. Zhou, Z., Mashuq, M., Sun, L.: Web content extraction through machine learning (2014)
8. Burget, R.: Layout based information extraction from html documents. In: ICDAR. (2007) 624–628
9. Burget, R., Rudolfova, I.: Web page element classification based on visual features. In: ACIIDS. (2009) 67–72
10. Sun, F., Song, D., Liao, L.: Dom based content extraction via text density. In: SIGIR. (2011) 245–254
11. Smith, R.: An overview of the tesseract ocr engine. In: ICDAR. (2007) 629–6332
12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations. **11** (2009) 10–18
13. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: ICCV. (2007) 1–8
14. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge. Computer Vision **88** (2010) 303–338