Content Extraction from Marketing Flyers

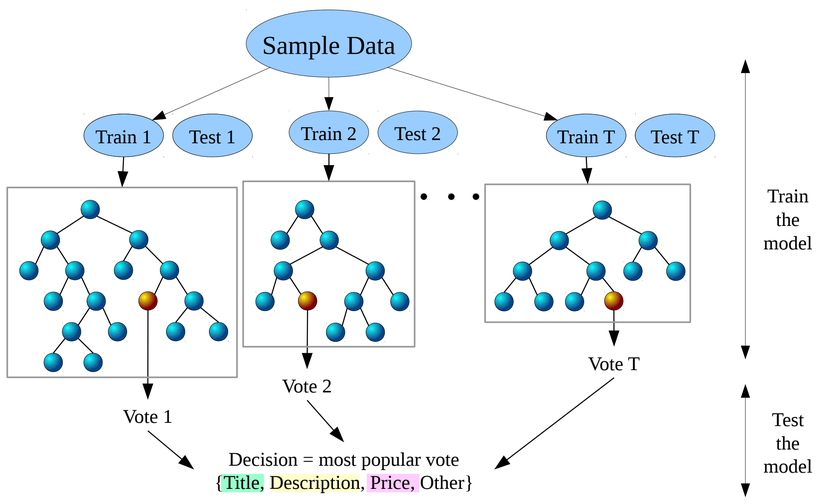
Ignazio Gallo, Alessandro Zamberletti and Lucia Noce
Università dell'Insubria, Varese, Italy
http://artelab.dicom.uninsubria.it/

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.

Random Forest

Divide training examples into T subsets



The Random Forest is an ensemble of Decision Trees each built on a random Subset of the input variables. The resulting models are highly accurate and have the added benefit of providing model error estimates and variable importance rating.

Tokenization

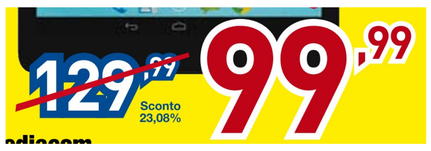
Textual features
is number
digits percentage
all upper case
only first upper case

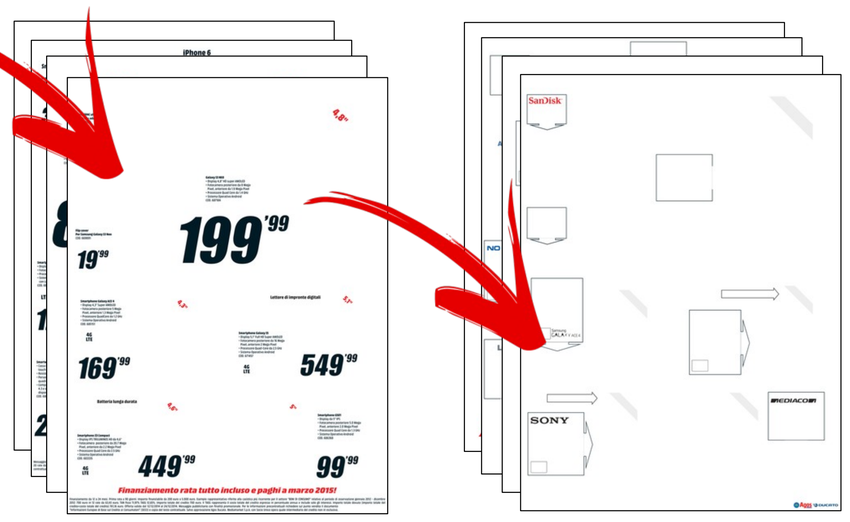
Visual features

Font
token font size
token angle
token position
token color
token font frequency
token color frequency
token markup color frequency
token font page frequency

Markup
token markup color frequency
token markup color

example





Sliding Window Pattern

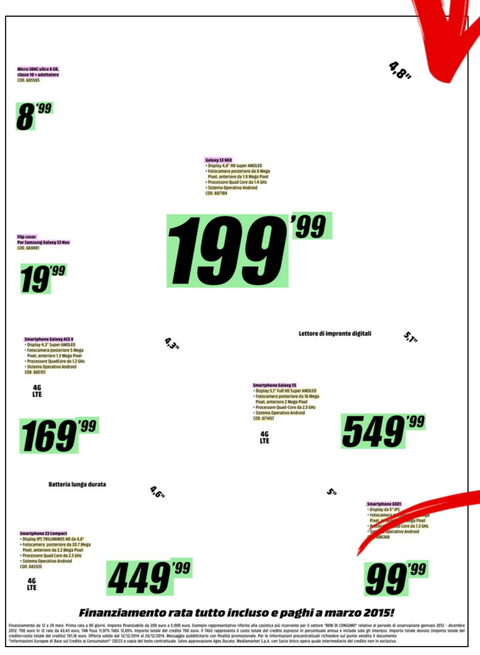
Left Token Features	Central Token Features	Right Token Features

Advanced Visual Features

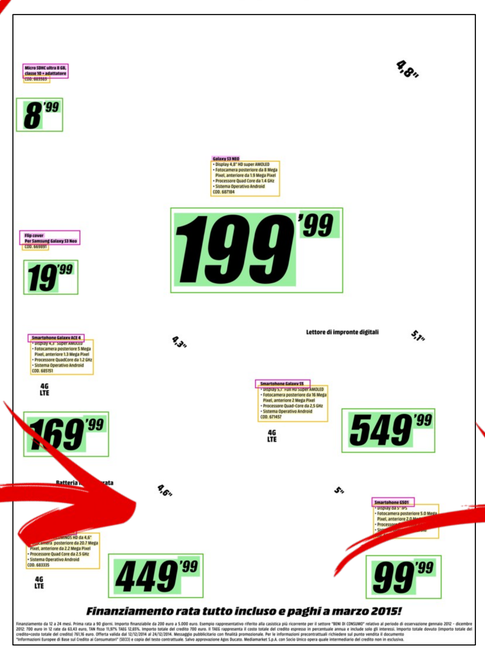
t is a token, ***p*** is a page and ***d*** is a document
f_t is the font of a token ***t***
$|p|$ is the total number of tokens in ***p***
n_{f_t} is the number of tokens having font f_t in ***p***
n_{c_t} is the number of tokens having font color c_t in ***p***
n_{m_t} is the number of tokens having markup color m_t in ***p***

Token **F**ont **F**requency $TFF_{t,p} = n_{f_t}/|p|$
Token **C**olor **F**requency $TCF_{t,p} = n_{c_t}/|p|$
Token **M**arkup **C**olor **F**requency $TMCF_{t,p} = n_{m_t}/|p|$
Font **P**age **F**requency $FPF_{t,d} = |\{p_i : f_t \in p_i\}|/|d|$

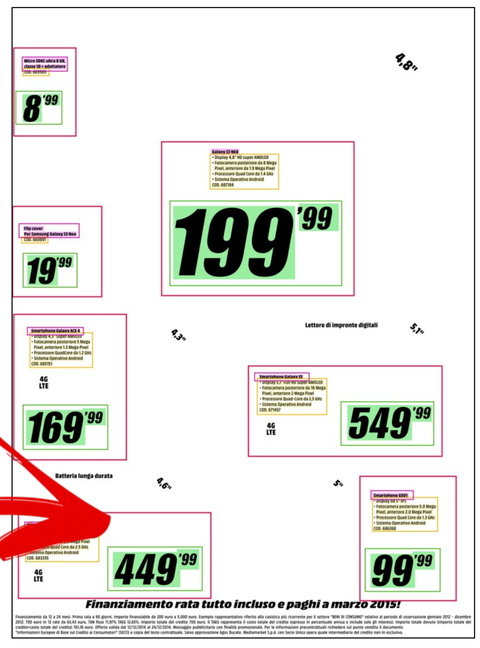
Token Classification



Token Aggregation



Offer Aggregation



Experiments

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.

Token classification confusion matrix

	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%

Token and offer aggregation

	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