Automatic Web Page Classification Using Visual Content for Subjective and Functional Variables

Nuno Goncalves(✉) and Antonio Videira

Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal
{nunogon,avideira}@isr.uc.pt

Abstract. Automatic classification of webpages has several applications in industry: digital marketing, search engines, content filtering and many more. Traditionally this classification has been done using only the textual information of webpages, which includes the html code, tags, title and more lately also the url. The aim of this paper is to prove that for some subjective variables, although very important to the applications mentioned, the visual information of webpages as they are rendered by the browser has extremely rich content for the classification task. The variables studied are the aesthetic value (whether pages are beautiful or ugly) and the design recency of them (whether pages are old fashioned or look modern). We then proved that automatic classifications that rely only on the visual *look and feel* can achieve very high accuracies. As we used several low-level and mid-level features and studied several criteria for selection and classification, our classifiers were able to improve one step further the stat of the art. Finally, we applied this framework to classify webpages in their topic (content aware) and also to classify whether pages are a blog or not (functional aware).

Keywords: Web page classification · Feature extraction · Feature selection · Machine learning · Blog classification · Aesthetic value · Design recency

1 Introduction

The automatic classification of documents, namely web pages, has been studied in the last years. The advantages of having a robot that is able to classify web pages in the most diverse variables, from those totally technical and objective, to those that are subjective, personal or even fuzzy, are significant and extremely current. This is the main task of web crawlers that browse all internet pages to classify them. Web page classification hence helps in focused crawling, assists in the development and expanding of web directories, helps in the analysis of specific web link topic, in the analysis of the content structure of the web, improves the quality of web search (e.g., categories view, ranking view), web content filtering, assisted web browsing and much more.

Most of the classification methods, however, rely only on the textual information of the web page. This methodology is highly convenient since it is fast

V. Monfort and K.-H. Krempels (Eds.): WEBIST 2014, LNBIP 226, pp. 279–294, 2015.
DOI: 10.1007/978-3-319-27030-2_18

and relatively content-rich for variables like topic or for segmenting their users for market purposes. Yet, textual information is very poor for some subjective variables usually connected to the so-called *look and feel* of a given web page. In fact, text and tags of web pages do not contain information about the aesthetic value of that page, neither about the recency of its design. There is, however, a rich content about these variables encoded in the visual information of pages, especially since nowadays the use of images and banners with embedded text is increasing and non-textual items are becoming unavoidable.

Furthermore, besides the usual classification of the topic of web pages, there is a less usual, but not less important, functional classification. For instance, since blogs are a significant channel for digital advertising, and since their content is changing in a daily base frequency, much higher than the crawler frequency of classification, it is important to build alternatives to classify these web pages in terms of functionality, more specifically whether they are blogs or not.

In this paper we are interested in the web page classification (WPC) in subjective variables, such as aesthetic value and design recency, relying only on visual features. We prove that the look and feel contain rich information for this classification and that it can be used to improve that classification task. We also prove that the visual features can be used to classify the page topic and is much effective for classifying blog functionality. This paper is an extended version of our work published at the 10th International Conference on Web Information Systems and Technologies (WEBIST-2014) [1].

2 Related Work

The text content that is directly located on the page is the most used feature. A WPC method presented by Selamat and Omatu [2] used a neural network with inputs based on the Principal Component Analysis and class profile-based features. By selecting the most regular words in each class and weighted them, and with several methods of classification, they were able to demonstrate an acceptable accuracy. Chen and Hsieh [3] proposed a WPC method using a SVM based on a weighted voting scheme. This method uses Latent semantic analysis to find relations between keywords and documents, and text features extracted from the web page content. Those two features are then sent to the SVM model for training and testing respectively. Then, based on the SVM output, a voting scheme is used to determine the category of the web page.

There are few studies of WPC using the visual content, because traditionally only text information is used, achieving reasonable accuracy. It has been, however, noticed [4] that the visual content can help in disambiguating the classification based only on this text content. Additionally, another factor in favor of using the visual content is the fact that subjective variables as design recency and aesthetic value cannot be studied using text content contained in the html code. These variables are increasing in importance due to web marketing strategies.

A WPC approach based on the visual information was implemented by Asirvatham et al. [5], where a number of visual features, as well as text features,

were used. They proposed a method for automatic categorization of web pages into a few broad categories based on the structure of the web documents and the images presented on it. Another approach was proposed by Kovacevic et al. [6], where a page is represented as a hierarchical structure - Visual Adjacency Multigraph, in which, nodes represent simple HTML objects, texts and images, while directed edges reflect spatial relations on the browser screen.

As mentioned previously, Boer et al. [4] has successfully classified web pages using only visual features. They classified pages in two binary variables: aesthetic value and design recency, achieving good accuracy. The authors also applied the same classification algorithm and methods to a multi-class categorization of the website topic and although the results obtained are reasonable, it was concluded that this classification is more difficult to perform.

3 Classification Process

In Fig. 1 it is possible to see the necessary steps to predict the class of new web pages. The algorithms were developed in C/C++ using the OpenCV library.

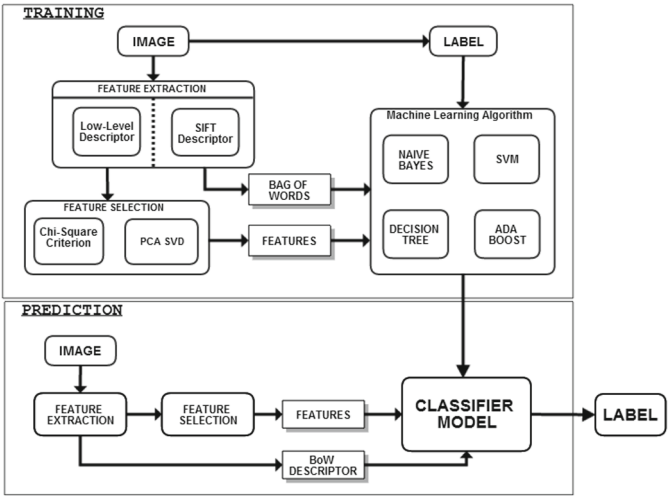

Fig. 1. Classification Process diagram.

3.1 Visual Feature Extraction

The concept of feature in computer vision and image processing refers to a piece of information which is relevant and distinctive. For each web page, different feature descriptors (feature vector) are computed. More details are presented in [1].

Low Level Descriptor. Visual descriptors are descriptions of visual features of the content of an image. These descriptors describe elementary characteristics such as shape, color, texture, motion, among others. To built this descriptor, with 166 feature-components, the following features were extracted from each image: color histogram, edge histogram, tamura features [7,8] and gabor features [9]. See more details in [1].

SIFT Descriptor Using Bag of Words Model. Keypoints are salient image patches that contain rich local information. The Scale Invariant Feature Transform was developed in 1999 by David Lowe. The SIFT features are one of the most popular local image features for general images, and was later refined by [10]. This approach transforms image data into scale-invariant coordinates of local features.

On the other hand, the bag-of-words (BoW) model [11] is a feature summarization technique that can be defined as follows. Given a training dataset D, that contains n images, where $D = \{d_1, d_2, ..., d_n\}$, where d are the extracted features, a specific algorithm is used to group D based on a fixed number of visual words W represented by $W = \{w_1, w_2, ..., w_v\}$, where v is the number of clusters. Then, it is possible to summarize the data in a $n \times v$ co-occurrence table of counts $N_{ij} = N(w_i, d_j)$, where $N(w_i, d_j)$ denotes how often the word w_i occurred in an image d_j.

To extract the BoW feature from images the following steps are required: (i) detect the SIFT keypoints, (ii) compute the local descriptors over those keypoints, (iii) quantize the descriptors into words to form the visual vocabulary, and (iv) to retrieve the BoW feature, find the occurrences in the image of each specific word in the vocabulary.

We use the SIFT and BoW implementations of OpenCV which outputs a 128-dimensional vector, training the classifier with different dictionary sizes: 100, 200 and 500 words.

3.2 Feature Selection

An important component of both supervised and unsupervised classification problems is feature selection - a technique that selects a subset of the original attributes by selecting the most relevant features. Two algorithms for applying feature selection are built. One is based on the Chi-Square Criterion, the other uses the Principal Components Analysis. In both methods a different percentage of the most relevant features is selected. To this work we used 1 %, 2 %, 5 %, 10 %, 20 % and 50 % of the total features.

As for the classifiers used, we opted for a wide variety of the most representative ones, by using their implementation of OpenCV. The classifiers used were then the Naïve Bayes, SVM, Decision Trees and AdaBoost.

4 Web Pages Database

In this work, different web page classification experiments are evaluated. There are three binary classifications and one multi-category classification. The three

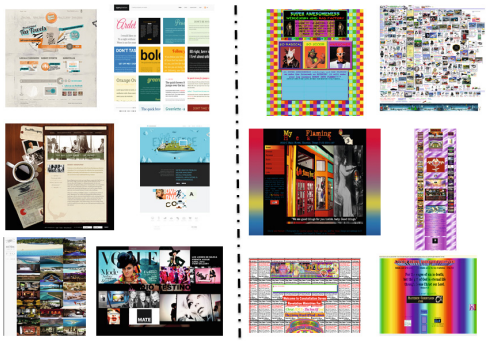

Fig. 2. An example of the web pages retrieved for the Aesthetic classification. In the left, there are 6 beautiful web pages, and in the right 6 ugly web pages.

binary classifications are: the aesthetic value of a web page, i.e., if a web page is beautiful or ugly (a measure that depends on the notion of aesthetic of each person), the design recency of a web page, i.e., trying to distinguish between old fashioned and new fashioned web pages and the blog functionality, i.e., whether a web page is a blog or not. The multi category classification involves classification on the web page topic.

Using the Fireshot plugin[1] for the Firefox web browser, allows to retrieve a screen shot of a web page and save it as a. PNG file. Different training sets of 30, 60 and 90 pages are built for each class of the classification experiment. For the blog classification we built a database with 800 blogs and 800 non blog pages. For each site we only retrieved the landing page which is generally the index page.

4.1 Aesthetic

The notion of aesthetic differs from person to person, because what can be beautiful for someone, can be ugly for another. That is why this classification depends of each classifier and it is a subjective classification. Nevertheless, there is a generic notion of the beautiful and of the ugly that is common to the individuals of a certain culture. We emphasize that this underlying notion of the aesthetic value is of extremely importance to marketing and psychological explorations.

In this classification experiment two classes are then defined: ugly and beautiful web pages. Notice that in Aesthetic, the important aspect is the visual design ("Look and Feel") of a web page, and not the quality of information or popularity of the page.

The ugly pages were downloaded from two articles [12,13] and their corresponding comment section, and also from the website World Worst Websites of the Year 2012 – 2005 [14]. The beautiful pages were retrieved, consulting a

[1] https://addons.mozilla.org/pt-pt/firefox/addon/fireshot/.

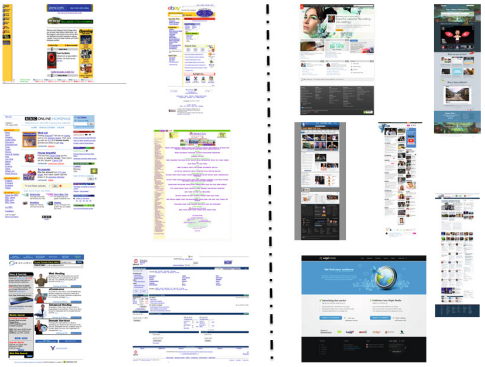

Fig. 3. An example of the web pages retrieved for the Recency classification. In the left, there are 6 old fashioned web pages from 1999, and in the right 6 new fashioned web pages from 2012.

design web log, listing the author's selection of the most beautiful web pages of 2008, 2009, 2010, 2011 and 2012 [15].

After analyzing the web pages retrieved (Fig. 2), it was possible to notice that, in general, an ugly web page don't transmit a clear message, uses too much powerful colors, lacks clarity and a consistent navigation. While, on the opposite side, a beautiful web page usually has an engaging picture, an easy navigation, the colors compliment each other and it is easy to find the information needed. Obviously these are some directives observed from the database and do not correspond to strict conclusions.

4.2 Design Recency

The objective of this classification is to be able to distinguish between old fashioned and new fashioned pages. The principal differences between these pages (Fig. 3) is that nowadays the web design of a page has firmly established itself as an irreplaceable component of every good marketing strategy. Recent pages usually have large background images, blended typography, colorful and flat graphics, that is, every design element brings relevant content to the user. In the past the use of GIFs, very large comprised text and blinding background were common in most sites.

The old web pages were retrieved consulting the article [16], that shows the most popular pages in 1999, and using the Internet Archive web site[2] allowed to retrieve the versions of those websites in that year. To retrieve the new pages, the Alexa[3] web page popularity rankings was used, selecting then the 2012 most popular pages.

[2] http://archive.org/web/web.php.
[3] http://www.alexa.com.

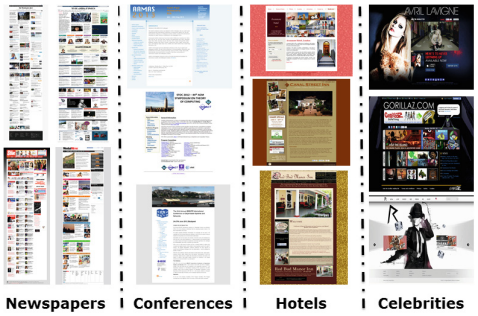

Fig. 4. Examples of web pages extracted for four web site topic classes.

Fig. 5. Examples of web pages extracted for the other four web site topic classes.

4.3 Web Page Topic

In this classification eight classes are defined: newspapers, hotels, celebrities, conferences, classified advertisements, social networks, gaming and video-sharing.

For the newspaper and celebrity classes, the http://Alexa.com was consulted, retrieving the most well-known and popular newspapers and celebrity sites. The celebrity sites also include popular fan sites. The conferences class consist in the homepages of the highest ranked Computer Science Conferences. And for the hotel class, different sites from bed-and-breakfast businesses are retrieved. The classes include different pages from different countries. The classified advertisements sites were extracted using also the http://Alexa.com, retrieving the most visited sites of classifieds of all world (sections devoted to jobs, housing, personals, for sale, items wanted, services, community, gigs and discussion forums). The video-sharing class and the gaming class (company gaming websites and popular gaming online websites), were extracted consulting the google search engine for the most popular sites in this type of websites. Social networks class consist in the major social networking websites homepages (e.g., websites that allow people to share interests, activities, backgrounds or real-life connections).

Fig. 6. An example of the blog pages retrieved.

A topic of a web site is a relevant area in the classification of web pages. Each topic has a relevant visual characteristic that distinguishes them, being possible to classify the web pages despite of their language or country. Looking at the pages retrieved (Figs. 4 and 5), it is possible to perceive a distinct visual characteristic in each class. The newspaper sites have a lot of text followed with images, while celebrity sites have more distinct colors and embedded videos. The conferences sites usually consist in a banner in the top of the page, and text information about the conference. Hotel sites have a more distinct background, with more photographs. Classifieds sites consist almost in blue hyperlinks with images or text, with a soft color background and banner. The body content of a video-sharing site consist in video thumbnails. The gaming sites have a distinct banner (an image or huge letters), with a color background and embedded videos. The social networks homepages, have a color pattern that is persistent.

4.4 Blog Classification

Classify a web page as being a blog or not can also be important for marketing and crawling reasons. This is why we added to this extended version of the paper [1] this study.

We thus built a database with 800 blogs and 800 non-blog pages. The blog pages were retrieved by random search in Google blogspot (http://blogspot.com). We retrieved 400 blogs written in Portuguese and 400 blogs written in English. For the non-blog pages we used random pages retrieved from the web index http://Alexa.com. See Fig. 6.

5 Results and Discussion

For the experiments, each classifier was evaluated with the low feature descriptor (containing 166 features), just the Color Histogram, Edge Histogram, Tamura Features, Gabor Features, and the descriptor containing the most relevant features selected by the methods of feature selection. Additionally the same data sets were used to train the classifiers with the SIFT descriptor using the bag of words model. The results for each classification task are shown in the next sections, as well as a comparison with the results of [4].

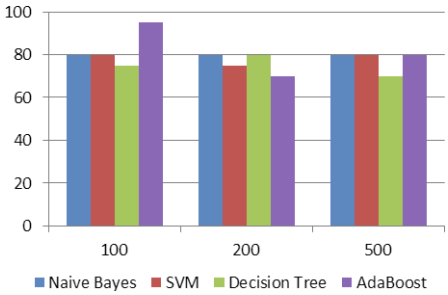

Fig. 7. SIFT Descriptor using BoW Model prediction results with different dictionary sizes (100, 200 and 500) for the Aesthetic Value.

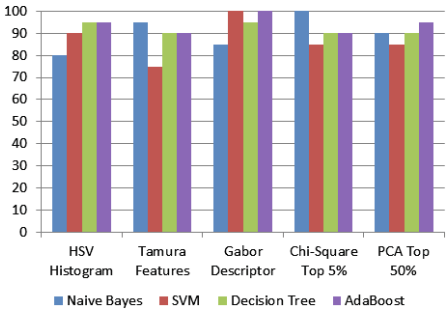

Fig. 8. Best prediction results for the Recency value for four different classifiers, using the low-level descriptor.

To test all methods after the training phase, new web pages were used to the prediction phase. Our results are based on the accuracy achieved by this prediction phase.

5.1 Aesthetic Value Results

In their experiments with the 166 features, [4] achieved an accuracy using the Naïve Bayes and a Decision Tree of 68 % and 80 % respectively. Using just the Simple Color Histogram and Edge Histogram they correctly classified 68 % and 70 % respectively for the Naïve Bayes, and 66 % and 53 % for the Decision Tree classifier.

For this experiment, Fig. 7 show the best rate prediction for our classifiers, when used the SIFT descriptor. Using different sizes for the dictionary, we obtained good result for each classifier. The best results for the Naïve Bayes, SVM and the Decision Tree was of 80 %, and for the AdaBoost we achieved a prediction accuracy of 85 %.

When trained the model using just the Color Histogram attributes, the results show an accuracy of 65 % for Naïve Bayes, 85 % in SVM, 70 % for the Decision Tree and 85 % using the AdaBoost when trained with 90 images for

each class. When we selected the top discriminative attributes to train the classifiers, the best results using the Chi-Squared method was when the classifiers were trained with the top 50 % attributes. The Naïve Bayes and SVM achieved an accuracy of 65 %, the Decision Tree 80 % and the AdaBoost an accuracy of 75 %. When trained with the top 20 % attributes by using the PCA method, the Naïve Bayes classifier achieved an accuracy of 75 %, the SVM classifier predicted 65 % of corrected pages, and finally, the Decision Tree and the AdaBoost classifiers both had an accuracy of 80 %.

All the classifiers showed a high prediction accuracy, with different features. Since most of the features chosen by the feature selection method are from the Color Histogram, it is possible to achieve a good prediction rate just by passing this simple descriptor. The SIFT descriptor give the best results, proving that the images from this two classes have distinctive keypoints.

5.2 Design Recency Results

In this experiment, [4] achieved an accuracy using the Naïve Bayes and a Decision Tree of 82 % and 85 % respectively. Using just the Simple Color Histogram the Naïve Bayes performed slightly worse than the baseline and the Decision Tree classifier sightly better. Using only the edge information, both models correctly classified 72 % and 78 % respectively for the Naïve Bayes and Decision Tree classifier.

Our best results for this experiment, using the low-level descriptor, are shown in Fig. 8. The Naïve Bayes, SVM and AdaBoost achieved an accuracy of 100 %, when the top 5 % attributes were selected using the Chi-Square method for the first one and the Gabor descriptor for the other two. The Decision Tree best accuracy (95 %), was when the PCA method selected the top 5 % attributes.

Relatively to the SIFT descriptor, all the classifiers obtained good accuracy. Notice that all the classifiers obtained an accuracy of 90 % when they used a dictionary size of 500. The best accuracy result achieved was for the Naïve Bayes with a 95 % rate of success, with a dictionary size of 200 words.

These results prove that the classifiers can learn just by using simple visual features. All the classifiers obtained good accuracy around 85 %, using the top 1 % attributes selected by both methods. Instead of using a more complex method like BoW, the use of simple visual features allows to decrease the computational cost for larger databases.

5.3 Web Page Topic Results

Experiment 1 - Four Classes. [4] define the following four classes for the topic: newspapers, hotel, celebrities and conference sites. The classification results obtained were the following: when all features are used, an accuracy of 54 % and 56 % for the Naïve Bayes and the Decision Tree respectively. Using the Color Histogram subset results in a much worse accuracy. Using only the Edge Histogram attributes, the Naïve Bayes predict with an accuracy of 58 %, whereas

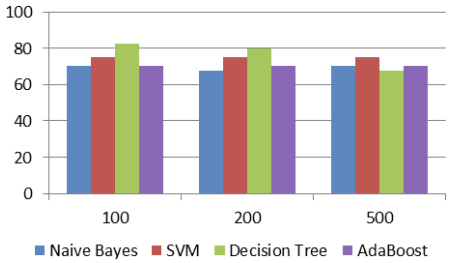

Fig. 9. SIFT Descriptor using BoW Model. Best prediction results with different dictionary sizes (100, 200 and 500). Experiments with 4 classes.

the Decision Tree predicts with an accuracy of 43 %. When they performed feature selection they show that the best predicting attributes are all from the Tamura and Gabor feature vectors. Using the top 10 attributes a prediction accuracy of 43 % for both classifiers was obtained.

Using the same low-level descriptor that they used, all our classifiers obtained better results. The Naïve Bayes achieved an accuracy of 62,5 % using the Tamura Features. The SVM and Decision Tree achieved an accuracy rate of 72,5 %, when used the selected top 20 % attributes using the PCA method and using the whole descriptor, respectively. While the AdaBoost classifier achieved an accuracy of 70 % using the PCA method selecting the top 50 % attributes.

Furthermore, the results showed in Fig. 9 are an improvement of the accuracy of approximately 22 % using the BoW model. Every classifier have an acceptable accuracy, where the best accuracy result is as high as 82,5 % for the Decision Tree using just 100 words to construct the dictionary. In fact, all the classifiers have accuracy higher than or equal to 70 % when used just 100 words in the dictionary.

Examining the results of the confusion matrices (Tables 1 and 2 corresponding to the best predictions of each classifier using the SIFT with BoW model (Fig. 9), it was verified, when analyzing the accuracy by class, that the Naïve Bayes, Decision Tree and AdaBoost perform much worse for the Hotel class. The Naïve Bayes and AdaBoost classifiers reports false positives for the Hotel class as Conference or Celebrity pages. While the Decision Tree returns false positives for Celebrities web pages as Hotel web pages, and vice versa. By his hand, the SVM classifiers perform much worse for the Celebrity web pages where most of the instances are erroneously classified as Hotel pages. Since the Newspapers and Conference classes have simpler designs, when compared with the other classes, they are easier to distinguish. On the other hand, it is harder to distinguish between more complex and sophisticated classes like Hotel and Celebrity.

Although the results obtained for this multi-class categorization are worse than those obtained for aesthetic value and design recency, generally good accuracy was obtained with best values usually near or above 80 %. Additionally, our results are better than those obtained by Boer et al. [4], mainly if SIFT with BoW is used.

Table 1. Confusion Matrix for 4 classes each with 10 web pages, for the best prediction result of the **Naïve Bayes** (table on the left) and **SVM** (table on the right) classifiers, using the SIFT descriptor.

		Actual - **Naïve Bayes**					Actual - **SVM**			
		Newsp.	Conf.	Celeb.	Hotel		Newsp.	Conf.	Celeb.	Hotel
Predicted	Newsp.	7	0	0	0	Newsp.	10	1	1	0
	Conf.	2	7	2	2	Conf.	0	8	1	0
	Celeb.	0	0	8	2	Celeb.	0	0	4	2
	Hotel	1	3	0	6	Hotel	0	1	4	8

Table 2. Confusion Matrix for 4 classes each with 10 web pages, for the best prediction result of the **Decision Tree** (table on the left) and **AdaBoost** (table on the right) classifiers, using the SIFT descriptor.

		Actual - **Naïve Bayes**					Actual - **SVM**			
		Newsp.	Conf.	Celeb.	Hotel		Newsp.	Conf.	Celeb.	Hotel
Predicted	Newsp.	10	0	1	0	Newsp.	10	0	1	0
	Conf.	0	9	0	1	Conf.	0	6	1	3
	Celeb.	0	1	7	2	Celeb.	0	1	8	3
	Hotel	0	0	2	7	Hotel	0	2	0	4

Experiment 2 - Eight Classes. Along with the four classes defined in the experiment 1, four additional classes were added to this classification: classified advertisements sites, gaming sites, social networks sites and video-sharing sites.

Using the low-level descriptor the Naïve Bayes had the best accuracy with 47,5 %, while the SVM achieved an accuracy of 41,25 % using the Tamura descriptor. The Decision Tree and AdaBoost classifiers had a poor performance, where the best accuracy was 37,5 % and 33,75 %, respectively. When we used the Chi-Square and PCA method to select the top attributes the classifiers performance didn't improve. We conclude that for this type of classification more complex features or a bigger database are necessary.

When we used the SIFT descriptor (Fig. 10) all the classifiers had a better accuracy relatively to the results obtained using the low-level descriptor. The SVM achieved an accuracy of 58,75 %, and the Naïve Bayes 63,75 %. The Decision Tree best accuracy was 48,75 %, while the AdaBoost only predict the correct class in 38,75 % of the predictions.

When examining the confusion matrices (Tables 3 and 4) of Naïve Bayes and SVM classifiers (which achieved accuracy over 50 % when using the SIFT descriptor), it is possible to verify that both classifiers have problems distinguishing celebrities web pages. The Naïve Bayes also struggles in identify Video-Sharing

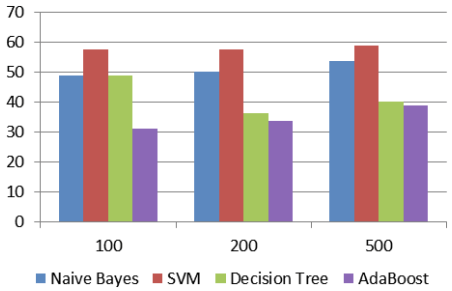

Fig. 10. SIFT Descriptor using BoW Model in experiment 2. Best prediction results with different dictionary sizes (100, 200 and 500).

Table 3. Confusion Matrix for 8 classes, by the **Naïve Bayes** classifier, using the SIFT descriptor.

		Actual							
		Newsp.	Conf.	Celeb.	Hotel	Classif.	Gaming	Social N.	Video
Predicted	Newsp.	9	0	1	1	3	1	0	4
	Conf.	1	5	0	0	1	0	0	0
	Celeb.	0	0	3	2	0	2	2	1
	Hotel	0	1	0	5	0	1	1	0
	Classif.	0	1	1	1	6	0	0	1
	Gaming	0	0	5	0	0	6	0	0
	Social N.	0	1	0	1	0	0	6	1
	Video	0	0	0	0	0	0	1	3

pages (only 3 correct predictions), while the SVM have troubles in identifying Social Networks web pages (only 2 correct predictions). The body of video-sharing web pages that consist mostly in video thumbnails are easily mistaken as newspapers web page (mostly images followed by text). In both methods some classifieds advertisements web pages are also predicted as newspapers (most classifieds advertisement websites use a simple color background with a lot of images). To overcome this drawbacks a bigger database is necessary.

5.4 Blog Classification Results

The results obtained for this binary classification are described by the confusion matrix Table 5. The results presented correspond to the best global accuracy, obtained using the Decision Tree classifier, achieving a global accuracy of 85 %. As can be observed the classification is much effective as expected, using only visual content.

Table 4. Confusion Matrix for 8 classes, by the **SVM** classifier, using the SIFT descriptor.

		Actual							
		Newsp.	Conf.	Celeb.	Hotel	Classif.	Gaming	Social N.	Video
Predicted	Newsp.	9	1	1	1	4	0	1	2
	Conf.	1	8	0	0	0	0	0	0
	Celeb.	0	0	4	2	0	3	2	1
	Hotel	0	0	0	7	0	1	1	1
	Classif.	0	0	1	0	6	1	0	0
	Gaming	0	0	4	0	0	5	2	0
	Social N.	0	1	0	0	0	0	2	0
	Video	0	0	0	0	0	0	2	6

Table 5. Confusion Matrix for Blog classification test (40 web pages per class), by the **Decision Tree** classifier, using the Low Level features (table on the left) and the SIFT feature descriptor (table on the right). Both descriptors achieved a global accuracy of 85 %.

		Actual - Low Level			Actual - SIFT	
		Blog	Non Blog		Blog	Non Blog
Predicted	Blog	34	6	Blog	34	6
	Non Blog	6	34	Non Blog	6	34

Furthermore, we emphasize that some Non-blog web pages that were classified as Blog pages are, actually, very close to the usual or typical blog. Figure 11 shows two examples of web pages classified as blogs that were in the non-blog database (false positives).

5.5 Discussion

The results show that based on aesthetic value and design recency, simple features such as color histogram and edges provide quite good results, where in some cases an accuracy of 100 % is achieved (average best accuracy of 85 %). It is also concluded that SIFT+BoW can also improve the accuracy at a considerably computational cost. For the topic classification, the use of a SIFT with BoW provide much better results too.

As expected when more website topics are added, the classification gets harder and the classifiers accuracy decreases to an average around 60 %. This indicates that even if the pages have visual characteristics that distinguishes

Fig. 11. False positives (non-blog pages classified as blog) that are, actually, very close to the typical blog.

them, they also have some attributes or characteristics in common. To overcome this setbacks a bigger database is necessary. Nevertheless, the aim of this work was to demonstrate that it is possible to classify web pages in different topics with reasonable accuracy and to prove that this visual content is very rich and can be successfully used to complement, not to substitute, the current classification by text crawlers. Notice too, that in the design of web pages, there is a growing tendency to include content in the images used, preventing text-based crawlers to get to this rich content (mainly in titles, separators and banners).

6 Conclusions

In this work we described an approach for the automatic web page classification by exploring the visual content "Look and feel" of web pages. The results obtained are quite encouraging, proving that the visual content of a web page should not be ignored, when performing classification.

In the future, in order to improve the classification accuracy we can also follow some additional paths. The integration of these visual features with other features of web pages can thus boost the accuracy in the classifiers. The analysis of the visual appearance of a web page can be combined with the well-established analysis based on text content, URL, the underlying HTML, or others. In this case, associate this visual features with the text content may give rise to a powerful classification system. Additionally, we also intend to mix the classification using visual features with a semantic analysis of them. We expect to improve the results by integrating the semantic content of a webpage image not only in the classification of the aesthetic or recency value but also for the classification of the topic. Another approach is the extraction of more sophisticated features that can analyze their dynamic elements (animated gifs, flash, advertisement content, and so on).

As for the applications of the visual classification of web pages, the methods studied may be applied to an advice system able to assist the design and rating of web sites to be applied to content filtering. In a research perspective, the fact that the aesthetic and design recency value are such a subjective measures, also make of great interest studies of the consumer profile for the field of digital marketing.

References

1. Videira, A., Goncalves, N.: Automatic web page classification using visual content. In: 10th International Conference on Web Information Systems and Technologies (WEBIST 2014) (2014)
2. Selamat, A., Omatu, S.: Web page feature selection and classification using neural networks. Inf. Sci. Inf. Comput. Sci. **158**, 69–88 (2004)
3. Chen, R.C., Hsieh, C.H.: Web page classification based on a support vector machine using a weighted vote schema. Expert. Syst. Appl. **2**(31), 427–435 (2006)
4. de Boer, V., van Someren, M., Lupascu, T.: Classifying web pages with visual features. In: 6th International Conference on Web Information Systems and Technologies (WEBIST 2010), pp. 245–252 (2010)
5. Asirvatham, A.P., Ravi, K.K.: Web page classification based on document structure. In: IEEE National Convention (2001)
6. Kovacevic, M., Diligenti, M., Gori, M., Milutinovic, V.: Visual adjacency multigraphs, a novel approach for a web page classification. In: Workshop on Statistical Approaches to Web Mining (SAWM), pp. 38–49 (2004)
7. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Trans. Syst. Man Cybern. **8**, 460–472 (1978)
8. Deselaers, T.: Features for Image Retrieval (thesis). RWTH Aachen University (2003)
9. Zhang, D., Wong, A., Indrawan, M., Lu, G.: Content-based image retrieval using Gabor texture features. In: IEEE Pacific-Rim Conference on Multimedia, University of Sydney, Australia (2000)
10. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **2**(60), 91–110 (2004)
11. Liu, J.: Image Retrieval based on Bag-of-Words model (2013). arXiv preprint arXiv:1304.5168
12. Andrade, L.: The worlds ugliest websites!!! (2009). http://www.nikibrown.com/designoblog/2009/03/03/theworlds-ugliest-websites. Acessed October 2009
13. Shuey, M.: 10-worst-websites-for-2013 (2013). http://www.globalwebfx.com/10-worst-websites-for-2013/
14. Flanders, V.: Worst Websites of the Year 2012–2005 (2012).http://www.webpagesthatsuck.com/worst-websites-of-the-year.html
15. Crazyleafdesign.com: Most beautiful and inspirational website designs (2013). http://www.crazyleafdesign.com/blog/
16. waxy.org: Den.net and the top 100 websites of 1999 (2010). http://waxy.org/2010/02/dennet_and_the_top_100_web-sites_of_1999/