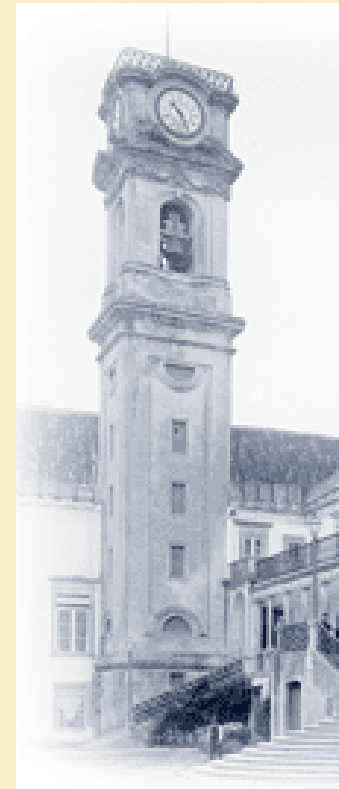


AUTOMATIC WEB PAGE CLASSIFICATION USING VISUAL CONTENT

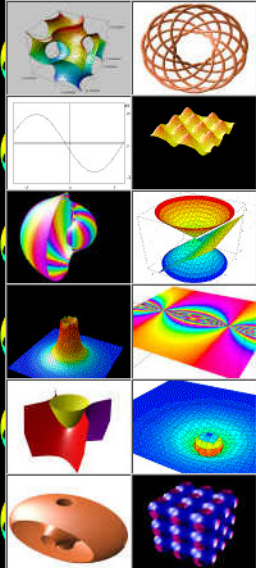


António Videira and Nuno Gonçalves

Institute for Systems and Robotics – University of Coimbra, Portugal
nunogon@isr.uc.pt

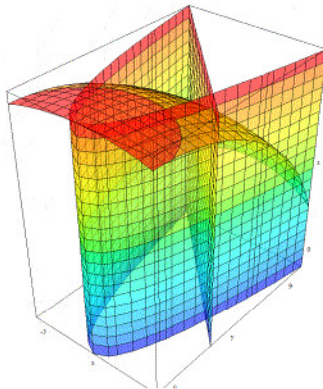
MOTIVATION

"Quality: Excellent. Value: Excellent." ... "DPGraph is one of the most exciting Windows-PC programs I've ever seen for creating beautiful, even stunning, mathematical graphics." Dr. Michael W. Ecker, Recreational & Educational Computing, and DPGraph runs under Wine on Linux, or under SoftWindows or Virtual PC on the Mac.



DPGraph: Dynamic Photorealistic 3D Graphing Software for Math and Physics Visualization

[Subscribe](#) | [List of Site Subscribers](#) | [Free Viewer](#)
[Latest news: 4 Jul 2008](#) | [Update to newest version: 4 Jul 2008](#)
[Math Art Gallery](#) | [Documentation](#) | [Links](#) | [Privacy](#) | [Contact](#)



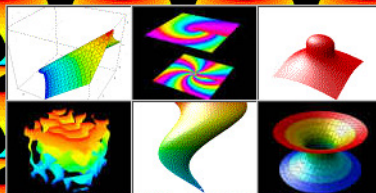
The world's most powerful software for math and physics visualization. Create beautiful, interactive, dynamic, photorealistic 2D, 3D, 4D, 5D, 6D, 7D and 8D graphs. So easy to use that even junior high and senior high students have had their graphs published. Includes hundreds of examples contributed by users from around the world.

Over two million mathematicians, physicists, teachers and students at over 1,000 colleges, universities and K-12 schools worldwide are already licensed. Comes with a free subscription to the [Flaming Thunder](#) programming language.

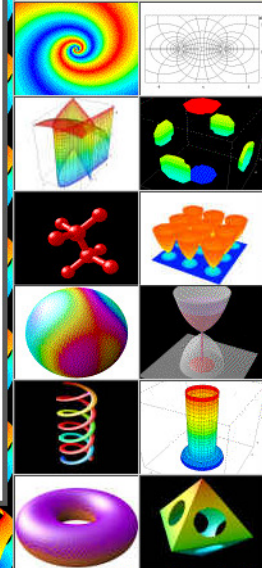
Optimized for the internet – search for DPGraph using your favorite search engine to find ways that people are using DPGraph for both classroom and distance learning. Used for pre-algebra, geometry, trigonometry and general physics, through multivariable calculus, field theory, quantum mechanics and gravitation.

Use time and color as extra dimensions (to create motion or encode momentum, for example). Use the scrollbar to vary parameters in realtime, to slice through graphs, or to vary transparency. Programmed entirely in assembly language for maximum speed.

Graph functions, equations, conic sections, planes, spheres, toruses, parametric curves and surfaces, implicit equalities and inequalities, volume intersections, volumes of integration, vector fields, surfaces of revolution, equipotential surfaces, and much more, in rectangular, polar, cylindrical, or spherical coordinates.



★★★★★ "You'll be dazzled ...", ZDNet



Copyright © 1997-2010 by David Parker. All rights reserved.

MOTIVATION



← 01/12 →



MOTIVATION

HELLO and welcome to
moviefone.com, from the people
who bring you 777-FILM



moviefone.com

A SERVICE OF AMERICA ONLINE



WEDNESDAY, OCTOBER 13, 1999, 4:03AM EDT

Showtimes & Tickets

...Plus everything else
you need to find a
movie

Word of Mouth

Instant feedback from
moviegoers. We'll even
tell you what critics say

Feature Presentations

Interviews, news, and
commentary from the
world of movies

visit our TRAILER PARK

The latest previews in
streaming video

sign up for MOVIEMAIL

Customized showtimes
of new releases via
e-mail

shop the MOVIEFONE STORE

Soundtracks, books,
videos, DVDs, and more

Dinner & A Movie?
Check out [Digital City](#)

© 1999 MovieFone, Inc.

[Privacy Policy](#)

SHOWTIMES & TICKETS

Enter your zip code to search:

and select a search option:

Search By Theater

Select a City or State

Select

International users [click here](#).

FREE Movie Tickets
for a year!
Random Hearts Contest!

Top five requested movies:

1. [Random Hearts](#)
2. [American Beauty](#)
3. [Three Kings](#)
4. [Superstar](#)
5. [Double Jeopardy](#)



[Click here for more info.](#)



STAR OF THE WEEK
Presenting Al Pacino



REALITY CHECK
A music teacher grades
Music of The Heart



FROM THE TRAILER PARK

Watch the most popular previews on moviefone.com



[Random Hearts](#)
Netshow 28.8 - 128
Real Video - 28.8 - 128
[Quicktime \(8.6MB\)](#)



[Superstar](#)
Netshow 28.8 - 128
Real Video - 28.8 - 128
[Quicktime \(11.4MB\)](#)



[Three Kings](#)
Netshow 28.8 - 128
Real Video - 28.8 - 128
[Quicktime \(6.8MB\)](#)



[American Beauty](#)
Netshow 28.8 - 128
Real Video - 28.8 - 128
[Quicktime \(9.8MB\)](#)

TODAY IN Word of Mouth

- [Top 10 Requested Movies](#): The most requested movies in your area. Updated hourly.
- [Top 10 Box Office](#): *Double Jeopardy* wins the box office battle.
- [First Feedback](#): Moviegoers rate *Superstar* and *Random Hearts*.
- [Pic of the Week](#): Mr. MovieFone sees... listens...tells. Find out what other moviegoers are buzzing about from Mr. MovieFone, and tell him what YOU [think](#).
- [Critics Digest](#): Critics dish on new releases.

TODAY IN Feature Presentations

- [Movies Opening This Week](#): *Random Hearts* and *Superstar* are coming to a theater near you.
- [Star of the Week](#): *Random Heart's* Harrison Ford.
- [Fall Movie Special](#): Upcoming must-sees, including *The Story of Us*.
- [Moviegoers' Journal](#): *Trick* - A deceptively good time.
- [Reality Check](#): A Gulf War vet grades *Three Kings*.
- [Movie Fan Site](#): *Molly's* Elisabeth Shue inspires devotion.
- [New Movie Sites](#): *Superstar* shines on the web.
- [Wish You Were Here](#): At the Miramax press junket.

MOTIVATION

neobux

[Ver Anúncios](#)

[Fórum](#)

[Registe-se](#)

[Iniciar Sessão](#)



Bem-vindo!
Agora pode multiplicar
os seus ganhos apenas
ao ver anúncios.

No NeoBux é pago apenas por ver
os sites dos nossos anunciantes.

[mais info](#)

[registre-se](#)



Quem pode ganhar

Membro

Como membro pode ganhar simplesmente por visualizar os anúncios que expomos.

- Lucro sem esforço
- Ganhe a partir de casa
- Anúncios diários garantidos
- Estatísticas detalhadas
- Oportunidades de expansão
- Comunidade dedicada
- AdPrize + Ofertas

Ontem

+ 10,022

Membros registados

Anunciante

Pode anunciar o seu website e aumentar as suas vendas e tráfego.

- Gestão avançada
- Milhões de potenciais clientes
- Filtro demográfico
- Forte proteção anti-fraude
- Estatísticas detalhadas
- Ao encontro dos seus objectivos
- Créditos no AdPrize

Ontem

+ 20,673,780

Visualizações de anúncios

Sobre nós

Somos especialistas em fornecer novas soluções de negócio num ambiente de ganho múltiplo.

- Ambiente seguro e estável
- Suporte profissional
- Serviços instantâneos
- Tráfego elevado
- Ideias inovadoras
- Orientados para o cliente

Ontem

+ \$107,072.33

Pago aos nossos membros



TESTED 18-JULY

© 2008-2013 NeoDev, Lda. Todos os direitos reservados.

[Termos de Serviço](#)

[Política de Privacidade](#)

[Ajuda](#)

MOTIVATION

The screenshot shows the homepage of The New York Times. At the top, the masthead reads "The New York Times" with the date "Monday, May 18, 2015" and the price "\$1.00 (US) \$1.00 (CA)". A prominent dark blue banner across the top of the content area offers "JUST 99¢ for 4 WEEKS OF A DIGITAL SUBSCRIPTION" with a "TRY IT NOW!" button and the text "The New York Times".

The main content area features several articles:

- Hackers From China Resume Their Attacks on U.S. Targets** by Steven Weber and Michael S. Roth. A column of the People's Liberation Army in China appears to have resumed its attack using different techniques, limiting removal of the same victims it has perpetrated in the past.
- Wells Fargo Suffers Losses in Florida Battle for Syrian City** by David Johnston. Fighting in the strategic city of Idlib was reaching a stalemate for the Lebanese militant group Hezbollah as it joined its Syrian allies in support of President Bashar al-Assad.
- Wells Dry, Fertile Plains Turn to Dust** by Steven Weber. Last year, fields that had 20,000 acres of pumpkins and melons were threatened by the summer drought of 2014.
- Explosion Shakes New England by 'Mop Treatment'** by Leah Green. Dying from the early harvest made of an invigoration in the Boston region American farmers' raised last year.

On the right side, there is a "The Opinion Page" section with articles like "Education, Violence and the Mayor's Race" and "The Great Divide: The 1 Percent Are Not 'Half the Problem'". Below this is a "MARKETS" section with a table showing stock market performance:

MARKETS	4/17/15	4/16/15	4/15/15
DAX	10,000.00	10,000.00	10,000.00
FTSE	6,000.00	6,000.00	6,000.00
NASDAQ	4,000.00	4,000.00	4,000.00
S&P 500	2,000.00	2,000.00	2,000.00

At the bottom right, there is a large promotional box for "JUST 99¢ for 4 WEEKS OF A DIGITAL SUBSCRIPTION" with a "TRY IT NOW!" button and an image of a smartphone displaying the app.

MOTIVATION

Newspaper



Hotel

Celebrities

MOTIVATION

M

ties

The screenshot shows the Britney Spears website with a large background image of her face. At the top, the name 'BRITNEY SPEARS' is displayed in a large, metallic font. To the right, there is a 'Member Login' section with fields for 'UserName' and 'Password', and buttons for 'Sign Up', 'Login', and a Facebook 'Log In' button. Below the name, a navigation bar includes links for 'TOUR', 'NEWS', 'BS ALERT', 'GALLERY', 'MEDIA', and 'SHOP'. A social media bar shows 'Follow @britneyspears' with a Twitter icon and '10M+ followers', along with icons for Facebook, YouTube, and other platforms. The main content area is divided into several sections: a 'SCREAM & SHOUT' album cover featuring Will.i.am and Britney Spears; a large red Google+ logo with the text 'Add Britney to Your Google+ Circles'; a 'BRITNEY SPEARS NEWS' section with a list of updates including 'Happy Mother's Day Britney!', 'Britney On The Cover Of Shape!', 'New Britney Song On The Way!', '"Scream & Shout" in Billboard's March Madness Finals', 'Vote For "Scream & Shout" On Billboard', 'The "Scream & Shout" Remix Video Is Here!', and 'Britney Tweets New S&S Remix Video Photo!'; and a section for 'Follow Britney On Instagram'.

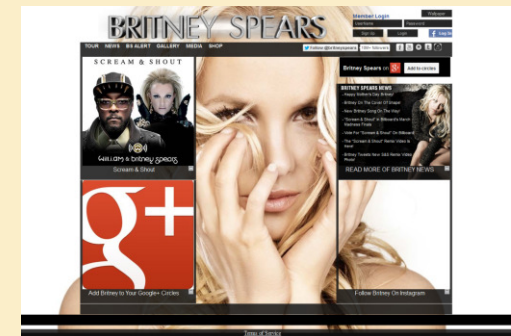
MOTIVATION

Newspaper



Hotel

Celebrities



MOTIVATION

New

Hackers From China Reveal Their Attacks on U.S. Targets

A screenshot of a news article snippet. The headline reads "Hackers From China Reveal Their Attacks on U.S. Targets". Below the headline, there is a small image of a person and some text, which is partially cut off.

SPONSORED BY Jeep

Obama seeks budget plan to avoid sequester
President Obama will propose short term budget plan Tuesday [Read Story](#)

Skier Lindsey Vonn airlifted after crash
Eric Willemssen, Associated Press
USATODAY

TOP NEWS MOST POPULAR

- Dell going private in \$24.4B leveraged ...
- Does your member of Congress own a g...
- Which colleges offer the best value?
- Feds file suit over S&P mortgage bond r...
- Judge denies delay in Zimmerman's trial
- Boy 'unharmed' after FBI rescue; captor ...
- Cantor: GOP agenda needs softer tone
- Tim McGraw enjoys newfound 'Freedom'
- Clydesdale in Budweiser ad gets name

HEADLINES Updated 11:18 a.m. EST

- NATION** Study calculates how much time to allow for traffic
- NFL** Joe Flacco bet millions on a winner: Himself
- BUSINESS** Unlikely stars emerge from Super Bowl ads

QUICK LINKS: Sports Scores, Crosswords, Sudoku, More Games

brities

RIGHT NOW

- 11:26 a.m. | Nation
Five-year-old hostage rescued, reunited with family: USA NOW video
- 11:26 a.m. | Tennis
Midland Challenger still going strong at 25
- 11:26 a.m. | Life
Jackman, Watts dish on Oscar luncheon
- 11:19 a.m. | Predators
Shootouts pile up pressure on Predators' Pekka Rinne
- 11:18 a.m. | World
Scientists reconstruct the face of 'Richard III'

Five-year-old hostage rescued, reunited with family: USA NOW video

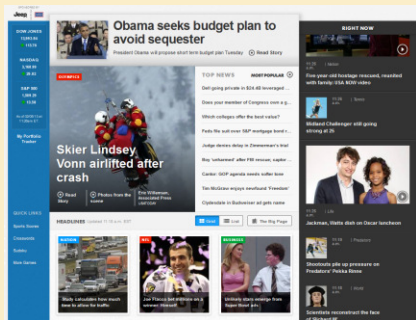
A screenshot of a news article snippet. The headline reads "Five-year-old hostage rescued, reunited with family: USA NOW video". Below the headline, there is a small image of a woman and some text, which is partially cut off.

MOTIVATION

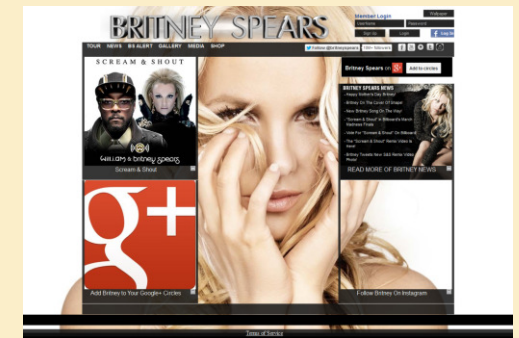
Newspaper



Hotel

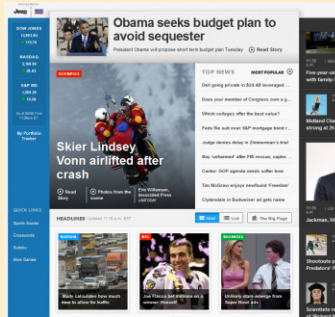


Celebrities



MOTIVATION

Newspaper



Home + Rooms & Rates + Photos + Location + Attractions + Contact us + **Book now**

Avonmore Hotel

57 Cartwright Gardens
London WC1H 9EL
Phone UK 0207 387 1939
Phone Outside UK 44 207 387 1939



Check Availability

To make an online booking select your arrival date and number of Nights.

Check-in Date
28 May 2013

Nights
1 Night

[Check Availability](#)

See our sister hotel
[Leigh House Hotel](#)

Avonmore Hotel, London

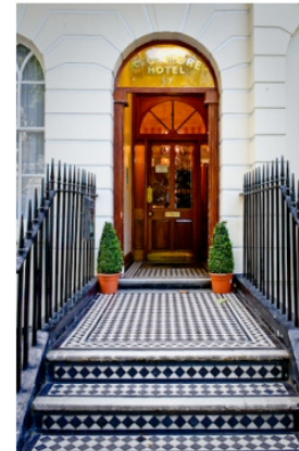
Avonmore Hotel with a combination of traditional and modern decor provides cosy and fresh rooms with Full English breakfast for tourist and business guests.

Located in the heart of London in Bloomsbury, the hotel is situated on a prime location of a Georgian Crescent overlooking tennis courts and gardens.

Avonmore Hotel is minutes from the British museum, British library, Covent Garden, Leicester Square and walking distance from Russel Square underground station and Kings Cross mainline station, the Avonmore Hotel offers the best of London.

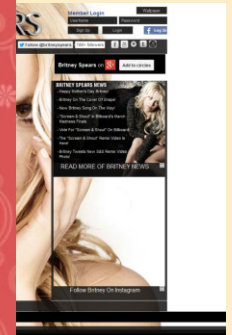
Its pleasant surroundings and central location, courtesy and comfortable atmosphere will ensure a pleasant stay in central London.

[Book Now!](#)



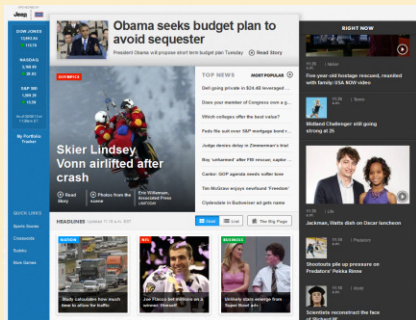
Avonmore Hotel - 57 Cartwright Gardens - London WC1H 9EL
Tel 0207 387 1939 - Fax 0207 388 3699
Email: bookings@avonmorehotel.net

ities

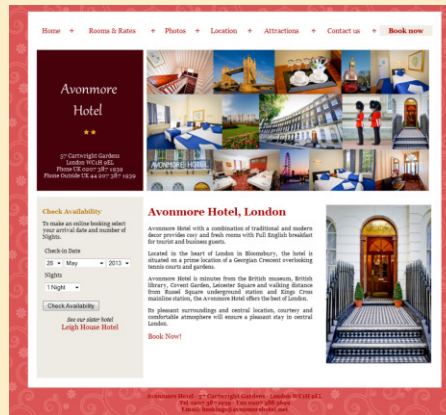


MOTIVATION

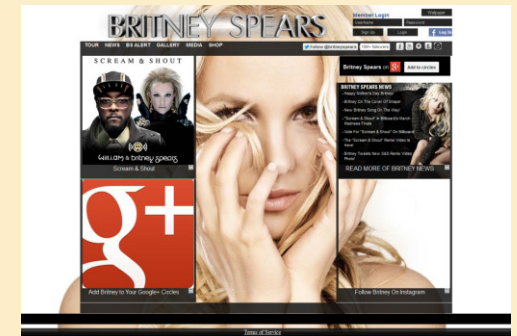
Newspaper



Hotel



Celebrities



OUTLINE

- ✗ Related work
- ✗ Problem statement
- ✗ Our approach:
 - + Feature extraction
 - + Feature selection
 - + Classifiers: training and testing
- ✗ Experiments
- ✗ Conclusions
- ✗ Future work

RELATED WORK

- ✗ Classification of web pages has been traditionally done using only text.
- ✗ Text is rich in semantic content.
- ✗ Classification is focused on the topic.
- ✗ Text sources:
 - + html code (tags: keywords, title, description, ...)
 - + html code (text in paragraphs or anchors)
 - + url (self url or page hyperlinks)

RELATED WORK

✗ Text only:

- + Neural networks + PCA [Selamat and Omatu, 2004]
- + SVM + semantic [Chen and Hsieh, 2006]
- + Web crawlers (Google, Yahoo, ...)

✗ Text + images:

- + Structure of page + images [Asirvatham and Ravi, 2001]
- + Visual Adjacency Multigraph (graphs) [Kovacevic et al., 2004].

✗ Visual only:

- + Several features + classifiers [Boer et al., 2010].

RELATED WORK

× Positives of using text:

- + very fast
- + easy to implement
- + rich content to classify the topic of the page.

× Negatives of using text:

- + it has a lot of irrelevant data
- + absolutely no information on some subjective variables

PROBLEM STATEMENT

- ✗ Classification of web pages
- ✗ Using only the visual content
- ✗ Subjective variables
 - + whether a page is beautiful or ugly – aesthetic value
 - + whether a page is old fashioned or has newer design – design recency
- ✗ Topic of the web page

VISUAL CONTENT

- × Positives of using visual content:
 - + rich content that we can not find in text
 - + can ben applied to subjective variables
 - + there is a growing trend to use images and banners in the design/layout of the page – these images contain visual text that is not caught by crawlers

- × Negatives of using visual content:
 - + slow

RICH CONTENT IN IMAGES OF TEXT

neobux

registe-se Iniciar Sessão

Bem-vindo!
Agora pode multiplicar os seus ganhos apenas ao ver anúncios.
No NeoBux é pago apenas por ver os sites dos nossos anunciantes.

mais info registe-se

Quem pode ganhar

Benefícios principais

Sobre nós
Somos especialistas em fornecer novas soluções de negócio num ambiente de trabalho múltiplo.

- Ambiente seguro e estável
- Suporte profissional
- Serviços instantâneos
- Tráfego elevado
- Ideias inovadoras
- Orientados para o cliente

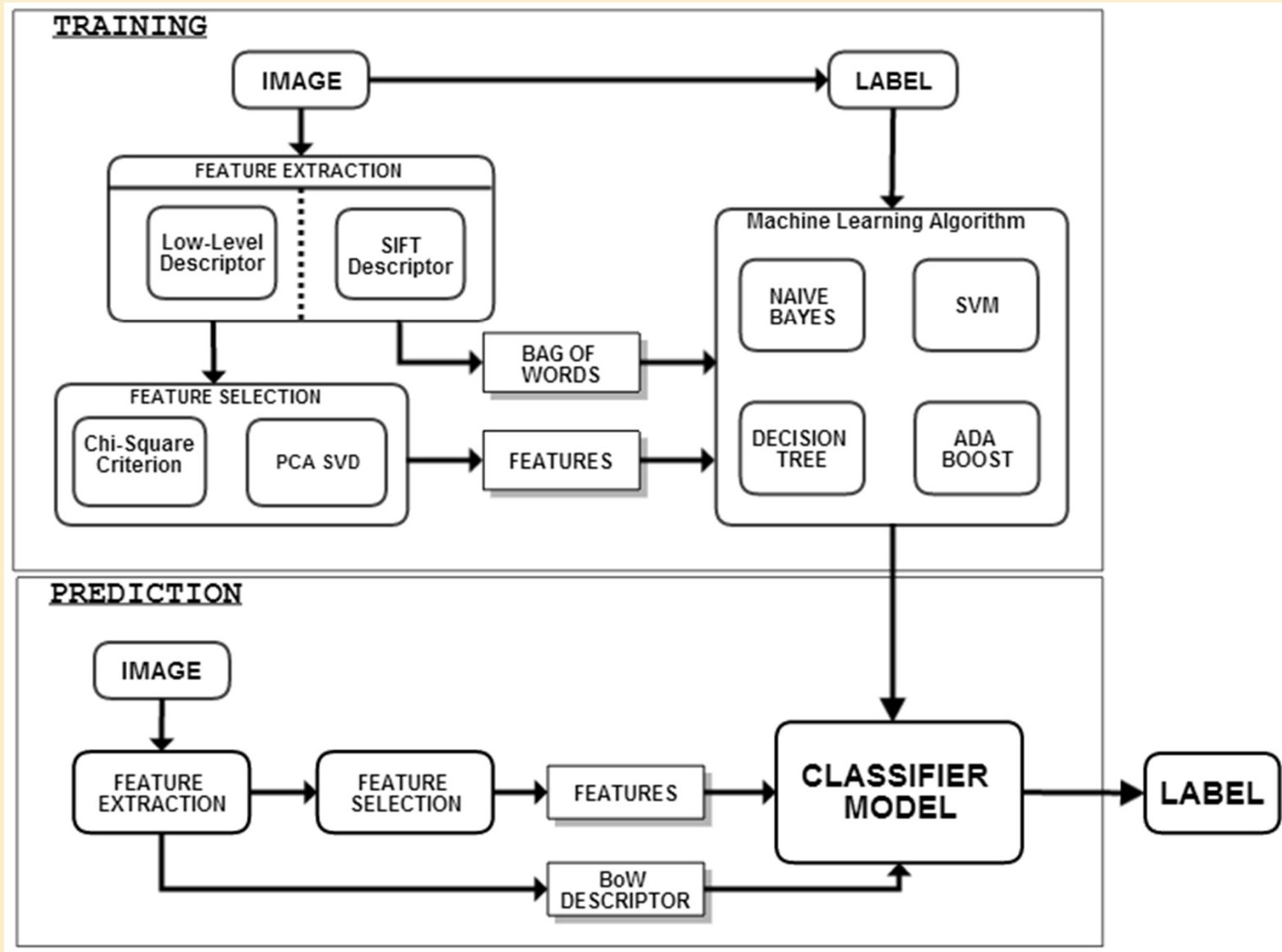
Membro
Como membro pode ganhar simplesmente por visualizar os anúncios que expomos.

Ontem
+ \$107,072.33
Pago aos nossos membros

- Lucro sem esforço
- Ganhe a partir de casa
- Anúncios diários garantidos
- Estatísticas detalhadas

Política de Privacidade Ajuda

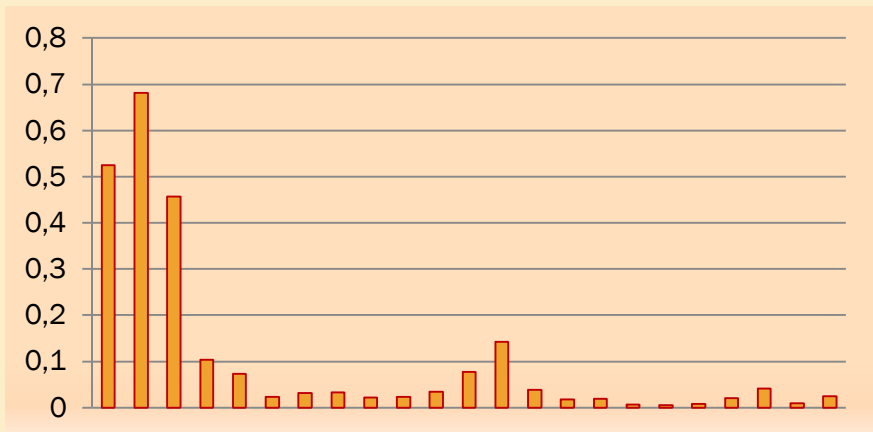
OUR APPROACH



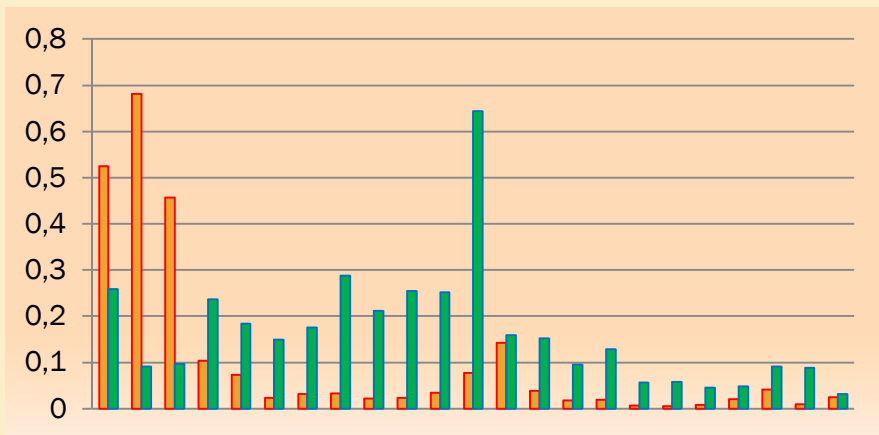
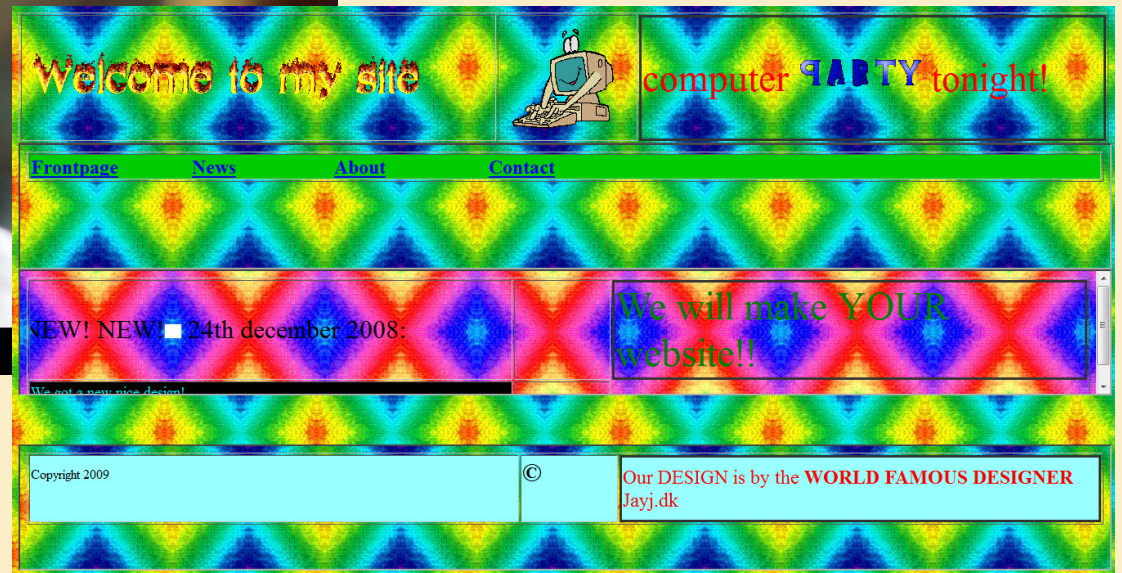
FEATURE VECTORS

- ✘ 1 – Low- level descriptor (166 dimensions)
 - + Color Histogram
 - + Edge Histogram
 - + Gabor
 - + Tamura
- ✘ 2 – Mid- level descriptor (100, 200 or 500 dimensions)
 - + SIFT
 - + Bag Of Words (BOW)

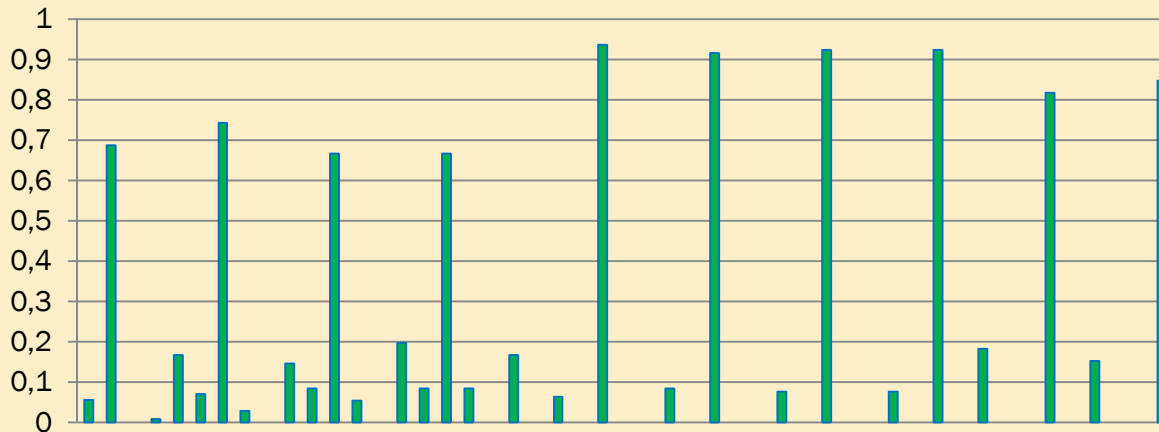
COLOR HISTOGRAM



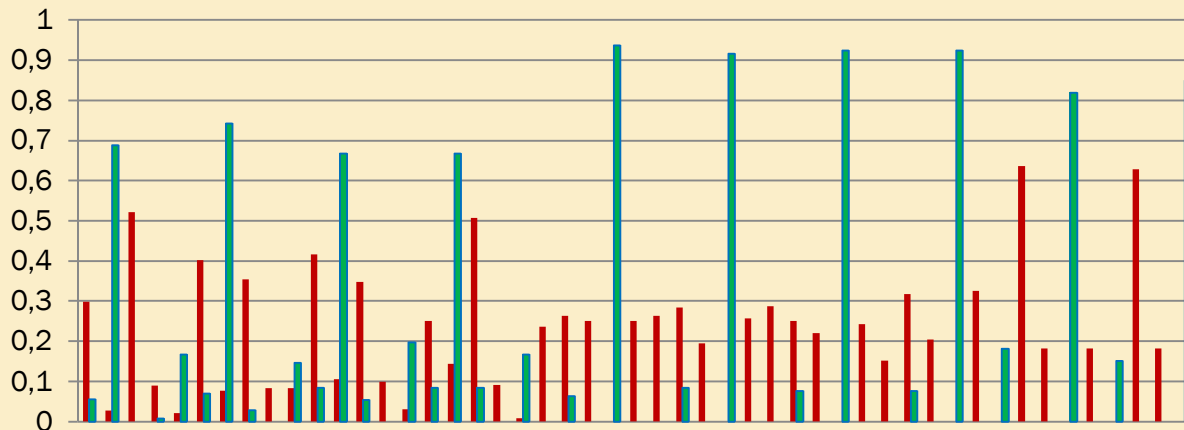
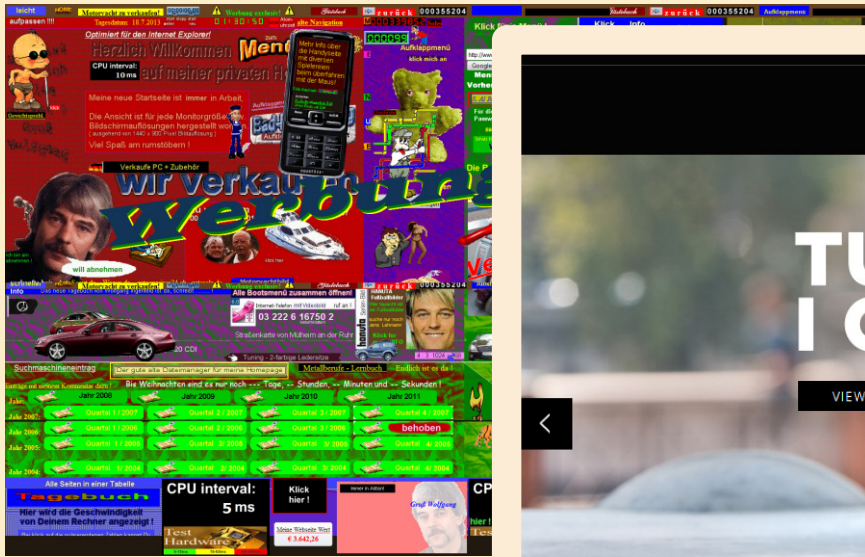
COLOR HISTOGRAM



EDGE HISTOGRAM



EDGE HISTOGRAM



GABOR FEATURES

www.irishwrecksonline.net
A GUIDE FOR DIVERS
 Click on a coastal grid for sub-maps & wreck diving details, or use the menu below

PRESS	List of all Wrecks by County
PRESS	List of dived Wrecks by Name
PRESS	Latest Updates
PRESS	Illustrations of Ship types
PRESS	Terminology used in this Site
PRESS	References used in this site
PRESS	Site Awards
PRESS	Sign irishwrecksonline's Guest Book
PRESS	View irishwrecksonline's Guest Book
PRESS	Want to join a scuba club in Ireland? <i>New!</i>

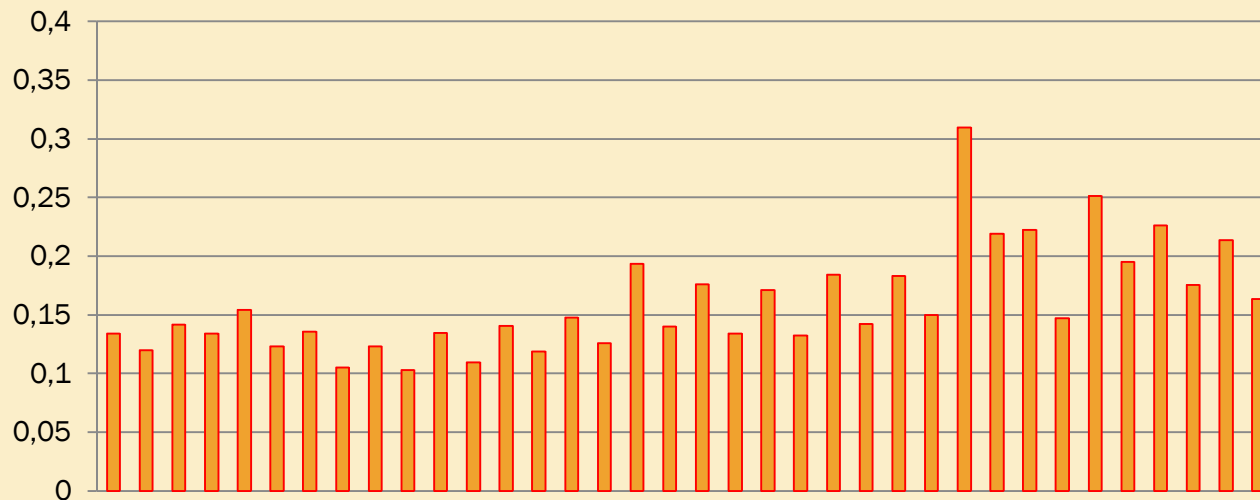
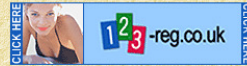
Search: irishwrecksonline ▾ for Search

Send errors, omissions, inclusions to
irishwrecksonline@yahoo.co.uk

© Copyright irishwrecksonline.net 1997 - 2012 - All external copyrights respectfully acknowledged



Do you have photos of any vessels listed on this site?
 Email me a copy if I can add them to the site. Full credit will be given. Thankyou.



GABOR FEATURES


www.irishwrecksonline.net
A GUIDE FOR DIVERS
 Click on a coastal grid for sub-maps & wreck diving details, or use the menu below

PRESS	List of all Wrecks by County
PRESS	List of dived Wrecks by Name
PRESS	Latest Updates
PRESS	Illustrations of Ship types
PRESS	Terminology used in this Site
PRESS	References used in this site
PRESS	Site Awards
PRESS	Sign irishwrecksonline's Guest Book
PRESS	View irishwrecksonline's Guest Book
PRESS	Want to join a scuba club in Ireland? New!


Search for Search

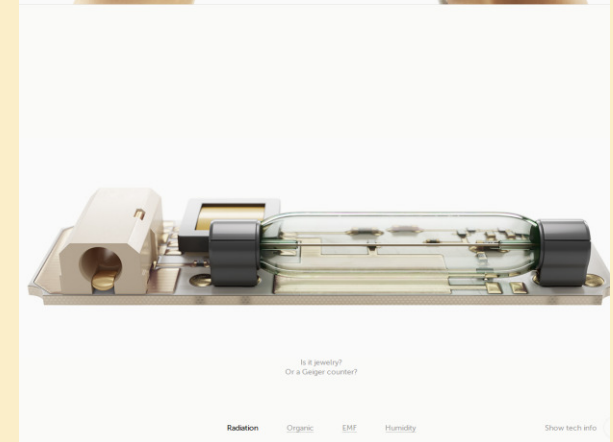
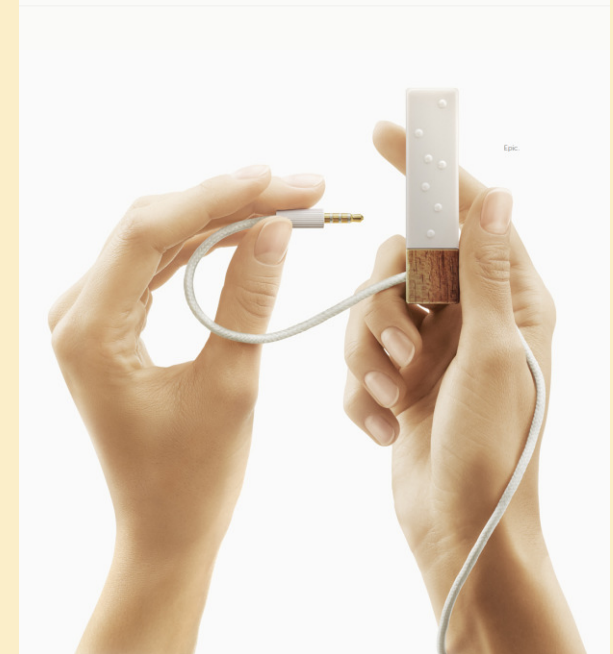
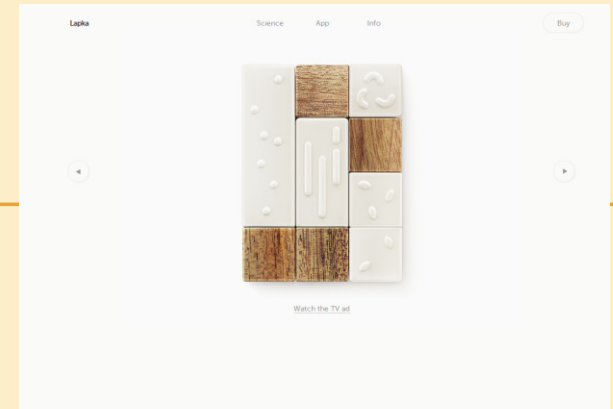
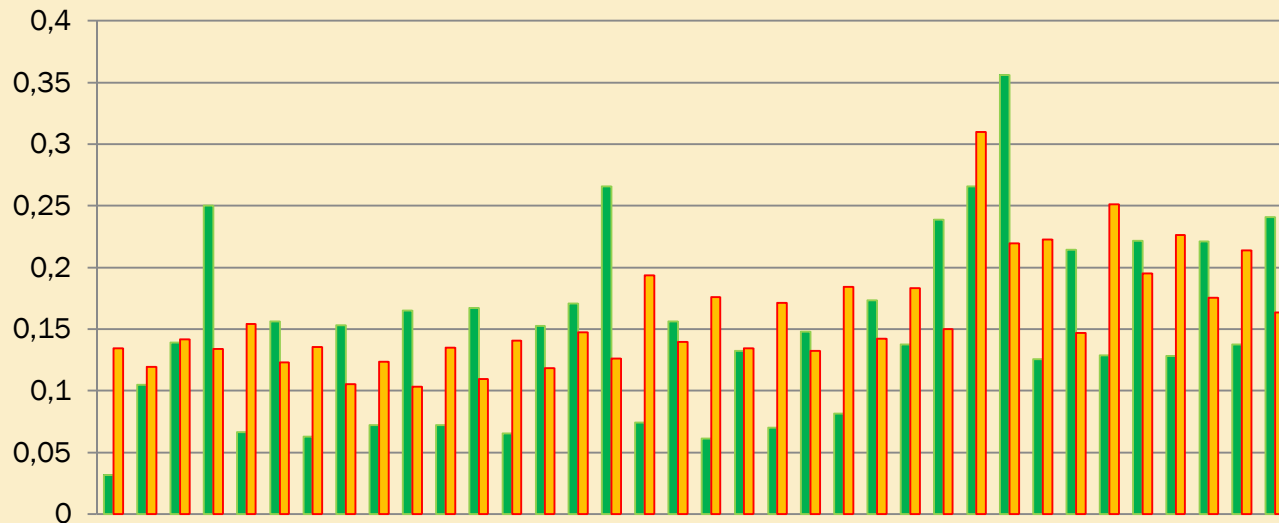
Send errors, omissions, inclusions to irishwrecksonline@yahoo.co.uk

© Copyright irishwrecksonline.net 1997 - 2012 - All external copyrights respectfully acknowledged



Do you have photos of any vessels listed on this site? Email me a copy if I can add them to the site. Full credit will be given. Thankyou.

CLICK HERE  BREK YOUR



TAMURA FEATURES

Lilywhite Bichons

AMERICAN KENNEL CLUB
AMERICAN KENNEL CLUB
BREEDER OF MERIT

Welcome to Lilywhite Bichons! Within our website, you'll find lots and lots of pictures of our beautiful and beloved Bichons Frise. We do hope you enjoy looking through our pictures and learning more about our Bichons and us. We invite you to sign our guestbook, thank you and enjoy your experience.

Our Lilywhite Bichons are featured in prestigious publications and calendars, including the American Kennel Club Gazette, Dog World November 2008 Cover picture of Billy Flynn, Dog Fancy, Barnes and Noble Bichon Frise Calendars, multiple Bichon Frise Calendars, and Checks Your Way.com (Bichon puppy checks).

2012 Calendars include:

- Barnes & Noble - 4 months within.
- Brown Trout Puppies - 2 months within.
- Willow Creek - 3 months within.
- Brown Trout - 1 month within.
- Brown Trout - Weekly - 4 months within.

2011 Calendars include:

- Avonside - 1 month within.
- Brown Trout - Cover - 1 month within.
- Brown Trout - Weekly - 1 month within.
- Silver Lining - 5 months within.
- Willow Creek - 4 months within.
- Barnes & Noble - 5 months within.

Past Calendar Publications - Others too numerous to mention.

Check our CALENDARS page to view these gorgeous pictures!

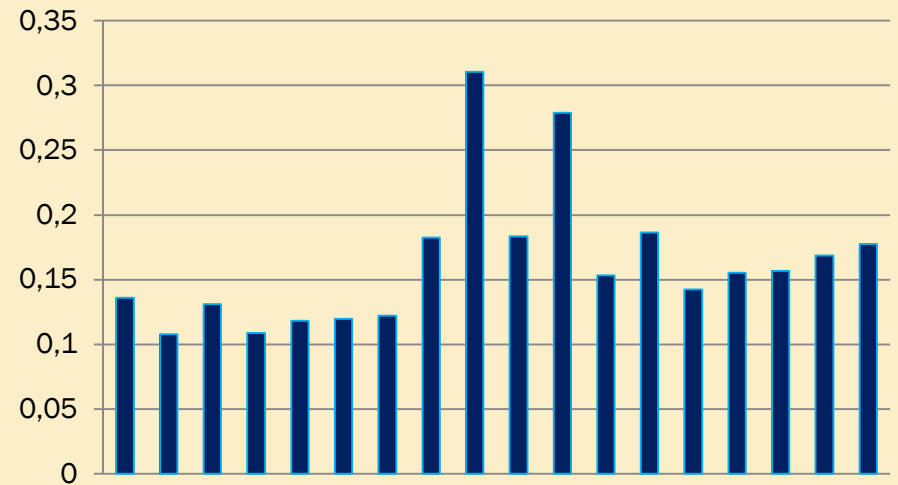
Home | About Us | Juno | Photos | Show Photos | Calendars
Puppies & Application | Prices | Standard | Links | Guest Book | Email

Contact Info
Janet and Jim Jamison
Asheville/Hendersonville Area in North Carolina
janjam98@aol.com

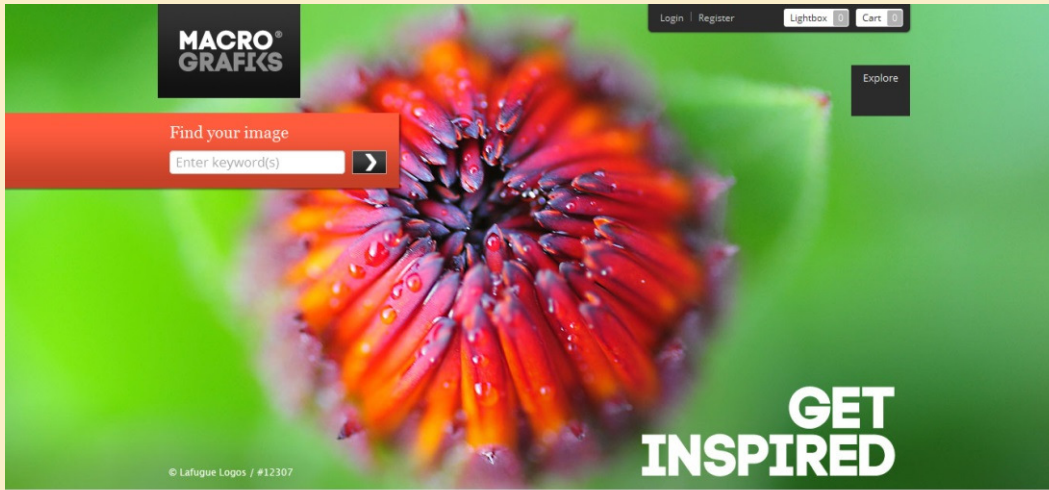
Home | About Us | Juno | Photos | Show Photos | Calendar |
Puppies | Prices | Bichon Standard | Links | Guestbook | Email

BOUTIQUE
VIEW SITE STATS

Maintained by Web Designs By Marie
Copyright © 2005-2012 Lilywhite Bichons. All rights reserved.



TAMURA FEATURES



Register now
Get 15 free credits (worth €15) as a new user



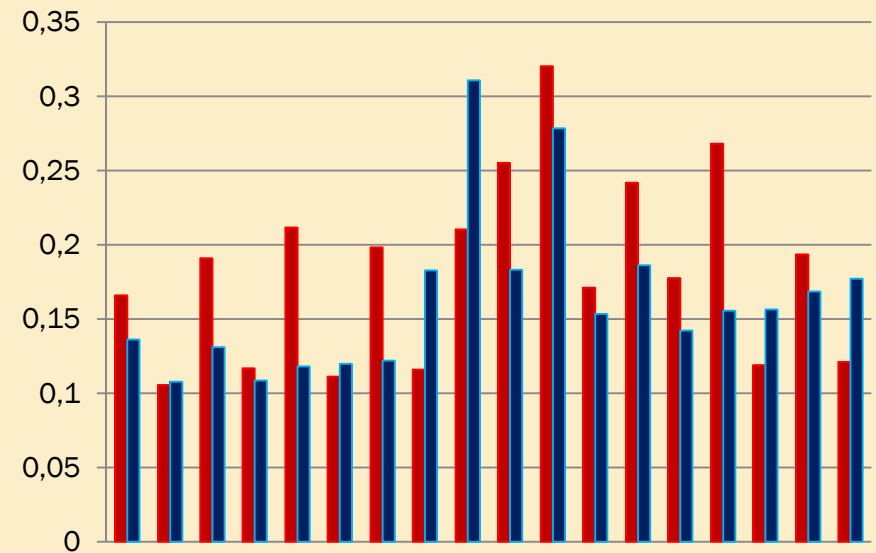
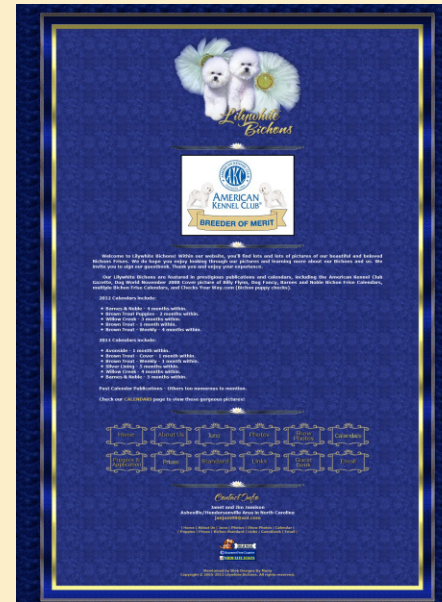
Five reasons
The best reasons to choose Macrografiks



60% Royalties
JOIN US
High royalties from the start

<p>Macrografiks</p> <ul style="list-style-type: none"> About Macrografiks Contact us Five reasons to trust us 	<p>Help</p> <ul style="list-style-type: none"> FAQs Licences Pricing & Formats 	<p>Photographers</p> <ul style="list-style-type: none"> Our contributors Become a contributor 	<p>Legal</p> <ul style="list-style-type: none"> Terms of Use Customer Agreement Privacy Policy 	<p>Follow us</p> <ul style="list-style-type: none"> Newsletter Facebook Pinterest Twitter
---	--	--	--	--

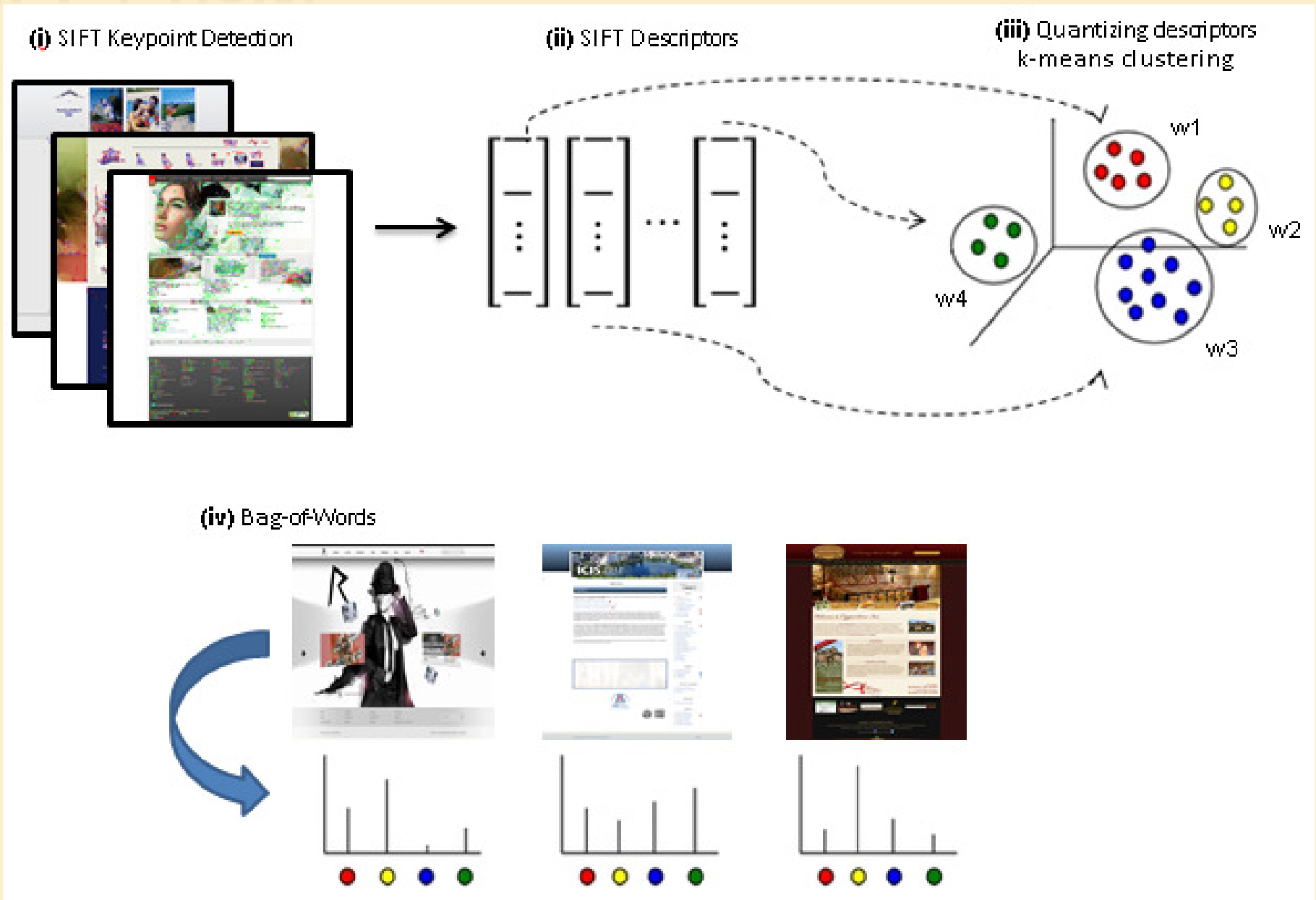
© 2013 Macrografiks. All rights reserved



SIFT + BoW

- ✗ SIFT features – David Lowe [1999, 2003]
- ✗ They are:
 - + scale and orientation invariant
 - + local descriptors
 - + slow to compute
 - + keypoint descriptor is an orientation histogram
- ✗ Bag of Words (BoW) – we use Jialu Liu [2013]
 - + All keypoint descriptors are used to build a dictionary by grouping SIFT features in “visual words”.
 - + A good trade-off between size and accuracy.

SIFT + BoW



SELECTION OF FEATURES

- ✘ Features are selected by their discriminative power.
- ✘ Chi Square criterion (χ^2)
 - + it relies on maximizing the matching between observed and expected frequencies.
- ✘ Principal Component Analysis (PCA)
 - + it relies on explaining the most variability of the sample

CLASSIFIERS

- ✘ Most representative classifiers were used in a supervised learning scheme:
 - + Naïve Bayes
 - + SVM
 - + Decision Tree
 - + AdaBoost

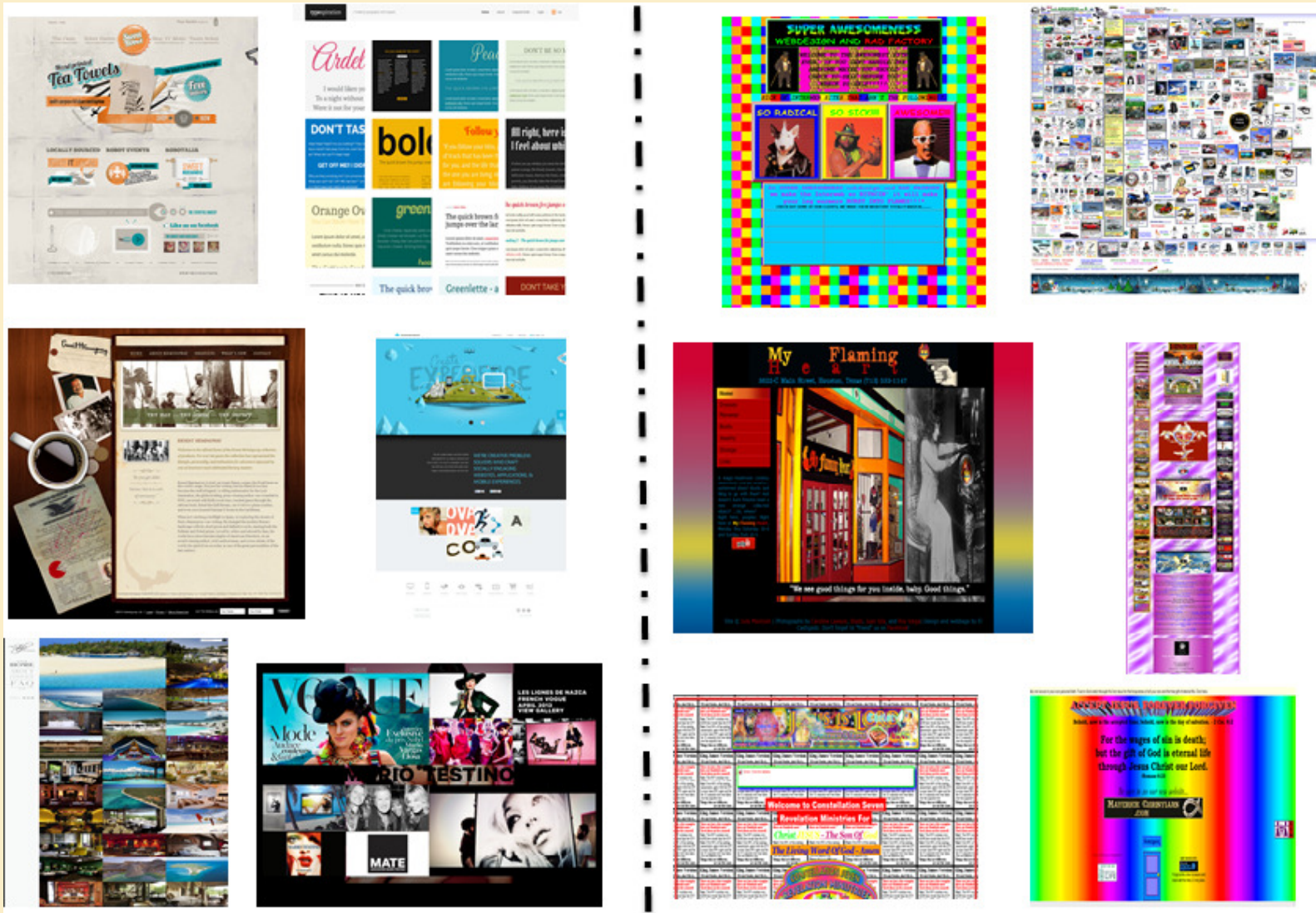
EXPERIMENTS

- ✗ 90 images for each class
- ✗ Images are renderings of landing web pages
- ✗ 2 binary classifications
 - + Aesthetic value (beautiful/ugly)
 - + Design recency (new/old fashioned)
- ✗ 1 multi-class classification
 - + Web page topic (4 classes and 8 classes)

AESTHETIC VALUE

- ✘ This is an inherently subjective variable
- ✘ Aesthetic notion is a valuable concept for marketing and psychology.
- ✘ Ugly pages are retrieved from blogs and public listings: [Andrade, 2009] and [Shuey, 2013].
- ✘ Beautiful pages are retrieved from designer blogs [Crazyleafdesign.com, 2013].

AESTHETIC VALUE EXAMPLES



DESIGN RECENCY

- ✘ This is also an inherently subjective variable, however, design recency can be more rationalized.
- ✘ Design notion of recency is a highly valuable concept for marketing.
- ✘ Pages were retrieved by consulting the most popular pages from today (2013) and for 1999, using internet archives and alexa.com.

DESIGN RECENCY



WEB PAGE TOPIC

- ✘ Topic is the main focus of web page classification.
- ✘ We started with 4 classes (inspired in [Boer et al., 2010]):
 - + Newspapers
 - + Hotels
 - + Celebrities
 - + Conferences

WEB PAGE TOPIC

- ✘ We then added 4 additional ones:
 - + Classified advertisements
 - + Social networks
 - + Gaming
 - + Video-sharing
- ✘ These 8 classes represent a big majority of all web pages. They are, however, not a systematic study on the topic.

WEB PAGE TOPIC EXAMPLES



Newspapers



Conferences



Hotels



Celebrities

WEB PAGE TOPIC EXAMPLES



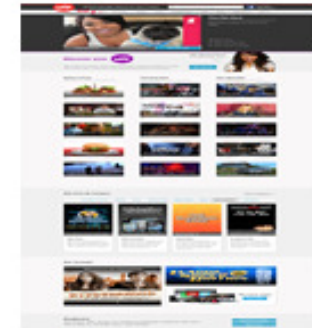
Classifieds



Gaming



Social Networks



Video Sharing

RESULTS - AESTHETIC VALUE

	Naïve Bayes	SVM	Decision Tree	AdaBoost
Color Histogram	65%	85%	70%	85%
Low-level (50%)	75%	65%	80%	80%
SIFT+BoW	80%	80%	75%	95%

- ✘ We observed that most of the features with higher discriminative power are collected from Color Histogram.

RESULTS - DESIGN RECENCY

	Naïve Bayes	SVM	Decision Tree	AdaBoost
Gabor only	85%	100%	95%	100%
Low-level (5%)	100%	85%	90%	90%
SIFT+BoW	90%	90%	90%	90%

- ✘ We observed that for design recency, Gabor features have the higher discriminative power. Gabor features are intrinsically related to texture and spatial frequency.

RESULTS - WEB PAGE TOPIC

4 CLASSES

	Naïve Bayes	SVM	Decision Tree	AdaBoost
Low-level	62,5%	72,5%	72,5%	70%
SIFT+BoW	70%	75%	82,5%	72,5%

- ✘ SIFT+BoW always improved results.
- ✘ Higher discriminative power is related with Tamura and Gabor features (texture, coarseness, frequency, ...).

RESULTS - WEB PAGE TOPIC

4 CLASSES

Predict\Actual	Newspapers	Conferences	Celebrities	Hotel
Newspapers	10	0	1	0
Conferences	0	9	0	1
Celebrities	0	1	7	2
Hotel	0	0	2	7

MAIN CONCLUSIONS

- ✘ Subjective variables are simpler to classify and features like color and edges have high discriminative power.
 - + Beautiful web pages tend to have more soft colors
 - + New fashioned web pages tend to have lower visual frequency and much lower density on information
- ✘ SIFT+BOW features are slower to compute and add only a small amount of discrimination.

MAIN CONCLUSIONS

- ✘ Web page topic is harder to classify using only visual content
- ✘ However, accuracies of 80% are generally achieved.
- ✘ SIFT+BoW features are able to enhance significantly the accuracy.
- ✘ A bigger database is, however, needed to achieve better results in classification of the topic.

CONCLUSION

- ✘ Visual content of web pages effectively has rich content to the task of classification in several variables: aesthetic value, design recency and web page topic.
- ✘ Classification using Visual content can be added to the traditional classification relying only on text to achieve a much powerful crawler, boosting its accuracy.

FUTURE DIRECTIONS

- ✘ Integrate visual content with text (html, url, ...) to achieve that powerful classifier.
- ✘ Explore the semantic content of the visual content, by segmenting the page and analyzing items separately (images, banners, flash, advertisement, ...).
- ✘ Analyze group differences of subjective variables and their implications on marketing (gender, age, culture, income rank, ...).

APPLICATIONS

- ✘ Building of an advice system to assist in the design of web pages.
- ✘ Marketing: targeting of consumers, campaign design, ...
- ✘ Content filtering and content scoring, including ranking of web sites based on user profile.
- ✘ Recommender systems for product web pages.
- ✘ Sentiment and emotional implications of web pages.

THANK YOU

- ✘ Some questions?
- ✘ I am searching for partners for H2020!
- ✘ Contact: nunogon@deec.uc.pt