https://doi.org/10.2352/ISSN.2470-1173.2017.13.IPAS-198

Brand detection framework in LG wavelet domain

Mangiatordi Federica, Bernardini Andrea, Pallotti Emiliano, Capodiferro Licia; Fondazione Ugo Bordoni; Roma, Italy

Abstract

This work proposes a method for an effective and quick monitoring of video contents produced by TV Broadcasters by means of a fully automatic system. The proposed system performs acquisition, recognition and classification of logos labeling video contents hosted by video-sharing platforms. This challenge is addressed in the Laguerre-Gauss wavelet domain; as soon as a logo is located, in any area of the video screen, a detection strategy, based on the analysis of local Fisher information of the selected logo region, is applied. A distance metric on the LG saliency maps, based nearest neighbor algorithm, is defined, to classify the logo in the relevant video portion. A preliminary test on a dataset of 300 heterogeneous videos, produced by several European Broadcasters, was performed, to verify the effectiveness of the proposed method. The experimental results proved the robustness of the implemented logo recognition and classification method, also for video content labeled with different logo sizes and shapes and for video content corrupted by geometric transformations and/or coding degradations.

Introduction

The World Wide Web has been witnessing an explosion of video content. We are living in an era of big and heterogeneous multimedia data.

"Digital technology and the Internet have created the most powerful instrument for the democratization of knowledge since the invention of moveable type for printing. They have introduced perfect fidelity and near zero-marginal costs in the reproduction of cultural works and an unprecedented capacity to distribute those works around the globe at instantaneous speeds and, again, near zero-marginal costs.."[1]

This paradigm stimulate the access and the creativity expression through downloading, editing and re publishing of any kind of video content. Several manipulations may occur and the drawback is the progressive lost of the relationship between the contents and their original producers (and right owners).

Unfortunately the fair use, a common practice in the context of textual contents where contents are reproduced but the source is credited, it is not yet utilized in the context of multimedia contents where the access and visualization have a direct economical value.

Video and multimedia contents are easy to copy and expensive to create.

On Internet there is no central entity to monitor and report the production of creative work, so it is not often possible to assess the ownership and the distribution rights.

Thus, it is a great target for illegal distribution, defined as piracy.

On one side copyright laws grant moral and economic rights to creator of a work on the other side downloading, editing and upload to online video platforms is not always original creative work. As an example a huge amount of videos on video platforms are unauthorized copies of television content.

Digital natives tend to think that all contents on internet are available for free, so trademark issues and counterfeit problems often get very little attention[2]. Therefore, a correct attribution of the economical value of video content became critical, in the balance between liberty of contents usage and content producers IP right protection. Youtube a world leader among video content portal has been dealing with internal tools for tackling IP infringement since its born and acquisition by Google in 2006 . Several strategies were applied, from content ID systems, to keywords analysis or content fingerprinting; none of them finally solving the problem and the volume of contents, for which IP rights were infringed, had an exponential growth[3]. Moreover new technical solutions have no backward compatibility (i.e. as content fingerprinting), so they apply to recent uploaded contents, leaving the enormous youtube archives left to itself.

A viable solution could be applied by means of the recognition of the content producers logos, that being a visible watermark, declares the intellectual rights ownerships. The issue of logo detection and recognition poses a wide number challenging questions to image processing experts, especially in the case of semi transparent logos that blend their luminance and color information with the hosting video background.

A variety of approach have been proposed in literature. Kleban et al. employed a multiresolution spatial pyramid mining technique for logo detection in natural scene [4]. Chang et al. developed a logo recognition method based on the extraction of the edge feature from the luminance variance map of the video segments [5]. Sahbi designed a variational framework to discriminate logos in static images by matching local features in [6]. Raluca Boia introduced the complete rank transform within a bag-of-words framework to increase the accuracy of their SIFT based brand detection system [7]. This paper addresses the logo recognition task in the Laguerre-Gauss wavelet domain. The local Fisher information of the detected logo region is exploited to select a robust set of key-points to be used in the classification process. Hence, a maximum likelihood functional is introduced to evaluate the similarity between two different LG saliency maps and recognize the TV Broadcaster logo following a nearest neighbor approach.

The rest of the work is organized as follows: the second section briefly presents the salient point extraction method based on the LG wavelet transform and the local Fisher information about position, scale and rotation [10]; the third section describes the framework for the proposed TV logo recognition system; the fourth section discusses the obtained results using a dataset of video clips collected from web; section fifth provides the conclusions.

Selection of salient points

Salient points are defined as points characterized by distinctiveness and invariance respect to geometric and radiometric distortions. These points are used in various multimedia understand-

ing applications as their neighborhood generally gathers discriminative information about the meaningful parts of the image. One fundamental requirement of any salient point detector is its stability respect to image transformations, like coding artifacts, resampling, photometric and geometric distortions. To achieve it we propose a wavelet based approach to compute the salient points of a candidate logo region. Specifically, it is considered the Harris-Fisher salient point detector that employs the Fisher information maps of local patterns to designate salient points according to specific criteria [10, 12].

To effectively summarize the key step of adopted method it is necessary to introduce some mathematical notations. Let be $f(x_1,x_2)$ an image on the real plane $\mathbb{R}^2$ and $f_w(\xi_1,\xi_2)$ its local pattern captured by a gaussian window $w(\cdot,\cdot)$ centered in the point (p_1,p_2),

$$\begin{aligned} f_w(\xi_1,\xi_2) &= w(\xi_1,\xi_2)\cdot f_0(\xi_1,\xi_2) \\ w(\xi_1,\xi_2) &= e^{-\frac{\xi_1^2+\xi_2^2}{s^2}} \\ f_0(\xi_1,\xi_2) &= f(x_1+p_1,x_2+p_2) \end{aligned} \tag{1}$$

Considering the Laguerre Gauss decomposition of image pattern

$$f_w(\xi_1,\xi_2) = f_w(rcos\theta, rsin\theta) = \sum_n \sum_k C_{n,k} g_{n,k}(r,\theta;s) \tag{2}$$

with $r=\sqrt{\xi_1^2+\xi_2^2}$ and $\theta = arctg\frac{\xi_2}{\xi_1}$, it is shown [10, 11] that the Fisher Information (FI) matrix of local pattern f_w about shift position, rotation and scale, can be obtained from the coefficients $A_{n,k}$. In eq(2) $g_{n,k}(r,\theta;s)$ represents the Laguerre Gauss function with angular order n, radial order k and scale factor s, defined as:.

$$\begin{aligned} g_{n,k}(r,\theta;s) &= h_n^{(k)}(r;s)\cdot e^{jn\theta} \\ h_n(r;s) &= \frac{(-1)^k}{\sqrt{k!\,(n+k)!}}\left(\frac{r}{s}\right)^n \frac{1}{s\sqrt{2\pi}} L_k^n\left[\left(\frac{r}{s}\right)^2\right] e^{-\frac{1}{2}\left(\frac{r}{s}\right)^2} \\ L_k^n(x) &= \sum_{i=0}^{k}(-1)^i \binom{n+k}{k-i}\frac{x^i}{i!} \end{aligned} \tag{3}$$

Specifically, the Fisher Information maps about the orientation J_ϕ and scale J_a can be computed by the following expressions:

$$J_\phi = \frac{4}{N_0}\sum_{n=1}^{\infty}\sum_{k=0}^{\infty} n^2 \mid C_{n,k} \mid^2 \tag{4}$$

$$\begin{aligned} J_a &= \frac{4}{N_0}\sum_{k=0}^{\infty}\mathscr{A}(k)\mid C_{0,k}\mid^2 \\ \mathscr{A}(k) &= \Gamma(\tfrac{3}{2})\Gamma(k-\tfrac{1}{2})\left\{\frac{1}{(k!)^2}+\frac{1}{8\pi[(k-1)!]^2}\right\}+\frac{\Gamma(2k+2)}{(k-1)!k!2^{2k+3}} \\ &+\frac{\Gamma(2k+1)}{2^{2k+2}}\left\{\frac{1}{(k-2)k!}+\frac{1}{[(k-1)!]^2}\right\}+\frac{\Gamma(2k)}{(k-2)!(k-1)!2^{2k+1}} \end{aligned} \tag{5}$$

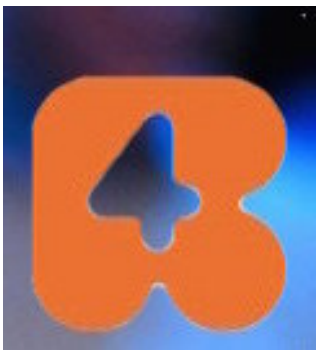
In addition, the FI of pattern about positional shift J_b can be related [10, 12] to the structure tensor $T_{\bigtriangledown f}$ by expressions in (6)

$$\begin{aligned} J_b &= \frac{1}{N_0} R_\phi T_{\bigtriangledown f} R_\phi^T \\ T_{\bigtriangledown f} &= \begin{bmatrix} \mathscr{E}_{f_x} & \mathscr{E}_{f_x f_y} \\ \mathscr{E}_{f_x f_y} & \mathscr{E}_{f_y} \end{bmatrix} \qquad R_\phi = \begin{bmatrix} cos\phi & sin\phi \\ -sin\phi & cos\phi \end{bmatrix} \\ \mathscr{E}_{f_x} &= \sum_{p=1}^{\infty}\sum_{q=1}^{p-1} B_{p-q,q}^2\left(\frac{p-q}{2s^2}\right) \quad \mathscr{E}_{f_y} = \sum_{p=1}^{\infty}\sum_{q=1}^{p} B_{p-q,q}^2\left(\frac{q}{2s^2}\right) \\ \mathscr{E}_{f_{xy}} &= \sum_{p=1}^{\infty}\sum_{q=1}^{p} B_{p-q+1,q-1}B_{p-q,q} \end{aligned} \tag{6}$$

where $B_{i,j}$ is the generic coefficient of Hermite expansion of local pattern $f_w(\xi_1,\xi_2)$ that are related to the coefficients C_{nk} [13] as follows:

$$\begin{aligned} C_{n,k} &= \frac{\pi}{\sqrt{(\mid n \mid +k)!k!}}\sum_{h=0}^{2k+n} j^{-h} g_{k+n,h}^{2k+n}\sqrt{(2k+n-h)!h!}B_{2k+n-h,h} \\ g_{l,k}^m &= \left(\frac{1}{2^m}\right)^{\frac{1}{2}}\sum_{q=max(0,h-l}^{min(m-l,h)}\binom{m-l}{m-l-q}\binom{l}{l-h+q} \end{aligned} \tag{7}$$

Once the FI maps are computed, the Harris-Fisher detector finds the salient points among the patterns characterized by local maxima of minimum eigenvalue of the gradient tensor (i.e. the FI about positional shift Jb), the relatively high invariance to rotation and the minimum dependence from scale variation. Denoting



(a) *Logo region*

(b) FI map J_a

(c) *FI map* J_ϕ

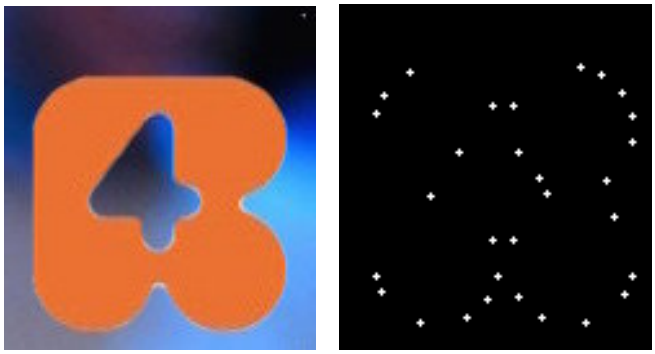
(d) *Minimum autovalue map of* J_b

Figure 1

with $H(p_1,p_2)$ the minimum eigenvalue of J_b the salient points are selected from the local maxima of $H(p_1,p_2)$ following the constraints:

$$\begin{aligned} J_a(p_1,p_2) &< TH_a \max_{(x_1,x_2)}[J_a(x_1,x_2)] \\ J_\phi(p_1,p_2) &> TH_\phi \max_{(x_1,x_2)}[J_\phi(x_1,x_2)] \end{aligned} \tag{8}$$

This approach increases the probability of selecting the same points at different scales and discards isolated points like spots, ensuring robustness against photometric differences, blur and compression artifacts.



(a) *Logo region* (b) *Saliency map*

Figure 2

Brand detection framework

The overview of the proposed system for video logo detection and classification is shown in Fig.3. The framework is constituted by two main cascading subsystems, denoted respectively as *Scraper* module and *Logo Detection and Recognition* (LDR) module.

The *Scraper* is an advanced tool to surf the web and collect video clips accordingly to specific keywords. This module extracts P*K frames from each identified video to feed the logo recognition component.

The *LDR* module processes each set of K frames downloaded from the same video clip and seeks segments and labels the region where the logo is located. After detecting the frame area $l(x_1,x_2)$ containing the TV brand, the LDR module analyzes this logo region in the Laguerre Gauss wavelet domain to extract its saliency map by means the Harris-Fisher detector presented in the previous section. Each salient point is associated with its LG features for constructing a set of N representative feature vectors that are processed by the logo classifier. To increase the robustness of the TV brand recognition system for intellectual property protection, the logo detection and classification method is repeated P times where P is the number of set of k frames extracted from the same video clip.

Scraper Module

The web *Scraper* module acts in the following two steps:

1 A *Scraper*, able to fully interact with a video-sharing website, which implements a headless browser (i.e. a windowsless browser for acting as web crawler) submits queries to the website, navigates the results and activates the video streaming;

2 Identification of K couples of video frames for the logo detection. In each couple to identify the first frame, we focalized on a random initial timing proportional to the video length minus an initial and final time segment to remove not relevant data (introductory content or credits). The second frame of the couple is selected with a delay corresponding to a fixed value (delay set at 60 seconds in the context of this research).

For the technical implementation of the Scraper we made usage of Selenium [8] an open source solution for automating web applications for testing purposes. Due the presence both of official channels for European broadcaster and not official channel to create an heterogeneous database of samples representative of European Broadcaster we manually selected 300 videos to be passed to the scraper modules.

Logo Detection and Recognition module

This module categorizes the video contents by performing three main steps:

- detection of logo region considering k frames at time;
- modelling of identified tv logo region in LG domain by the construction of a set of key-points;
- classification of identified logo region on the basis of the set of its feature vectors.

In the rest of this section, it is detailed each of the listed steps.

Logo detection

The method for the TV logo detection and segmentation is a modified version of the Yan technique[9] based on frame differencing. It exploits K video frames obtained by the random temporal sampling of the video contents. Given the sequence $\{f_j(x_1,x_2), j \in [1,K]\}$ of downloaded frames, it is computed the enhanced image $v(x_1,x_2)$ as written in eq. (9)

$$V(x_1,x_2) = \sqrt[\frac{K}{2}]{\prod_{j=1..\frac{K}{2}} \left[\sqrt{f_j * f_{j+3}} - \alpha(|\, f_j - f_{j+3}\,|)\right]} \tag{9}$$

Then the maximum difference between frames is derived by solving the expression:

$$d(x_1,x_2) = \max_{j=1..k} |\, v(x_1,x_2) - f_j(x_1,x_2)\,| \tag{10}$$

It is expected that the frame area overlapped by the TV logo is characterized by small temporal variation of luminance and color, consequently the distance $d(x_1,x_2)$ will assume extremely small values. This suggests to identify the bounding box of the logo region as the smallest rectangular area with the greatest density of pixels associated with $d(x_1,x_2) < \varepsilon$. The threshold ε is set dynamically as 0.5th percentile value of $d(x_1,x_2)$.

Once localized the logo bounding box, this is used to filter out the pixels belonging to the movable objects and background from the frame $f_j(x_1,x_2)$ and the enhanced image $v(x_1,x_2)$. Considering the enhanced version $v(x_1,x_2)$ the logo recognition task is simplified when the video frames are overlapped by a semi-transparent TV logo. In this case, the image $v(x_1,x_2)$ presents a more contrasted logo region and an attenuated video background.

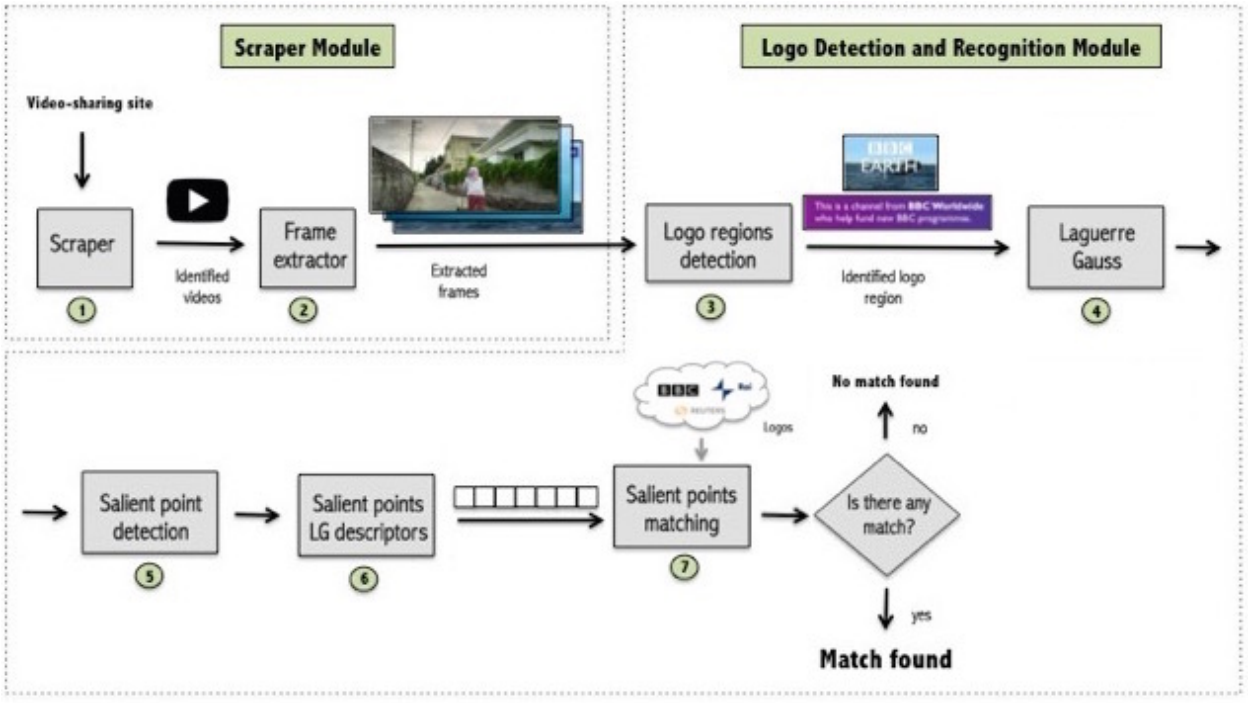


Figure 3: Tv logos monitoring framework

Logo modelling in LG domain

To recognize the logo class and label each logo region, it is assumed to model each identified logo region with a set of N feature vectors that gather the local information associated with the neighborhood of the salient points.
The main idea at the basis of this approach is that if two images contain the same logo, their salient points are roughly the same so as their local information. Hence, the image corresponding to the identified logo area is processed with a bank of LG filter [10] to compute the coefficients $C_{n,k}$ of its Laguerre Gauss decomposition and subsequently compute N representative salient points using the Harris Fisher detector. This leads to the construction of a saliency map for each logo region which can be described by a bag of N visual words constituted by the LG coefficients $C_{n,k}$ of salient points ($n = 0..4, k = 0..4$).

Classification

Given the logo templates of the main European TV broadcasters, the set of corresponding key points can be considered as the distinctive set of visual words for the logo recognition. Evaluating the minimum distance between the visual words of a given logo region and those of logo templates, allows to infer the correct TV label to assign to each video content.
Specifically, given the set of salient points $\{P_i = (p_1, p_2)_i; i = 1..N\}$ and $\{Q_j = (q_1, q_2)_j; j = 1..N\}$, respectively of logo template $l_{TV}(x_1, y_1)$ and logo region $f(x_1, y_1)$, the distance between them is defined as Gauss-Laguerre Log-Likelihood the functional [11] of the saliency map of the logo template given the saliency map of the identified logo region:

$$d = \sum_{P_i} ln\Lambda[l_{TV}(P_i); \hat{a}, \hat{\phi} | Q_j] \tag{11}$$

It is shown [11] that the Log-likelihood functional can be computed by the approximating expression (12):

$$\begin{aligned} & ln\Lambda[l_{TV}(P_i); a, \phi | Q_j] = \\ & = -\frac{2}{N_0} \cdot \min_{Q_j} \sum_{n,k} | D_{n,k}(P_i) - B(\hat{a}; n, k) C_{n,k} e^{-jn\hat{\phi}} |^2 \\ & B(a; n, k) = a^{-n-2k} \end{aligned} \tag{12}$$

where $D_{n,k}$ and $C_{n,k}$ are the LG coefficients associated with the salient points P_i and Q_j respectively. The parameter $\hat{a}$ and $\hat{\phi}$ eventually represent the scale and rotation factors of the logo region respect to logo template. Hence, the logo classifier computes all distance between the selected logo region $f(x_1, y_1)$ and the TV logos $l_{TV_r}(x_1, y_1)$ to find the nearest neighbour template and determine which logo category to associate with the logo region.

Experimental Results

The experiments were carried out on a dataset of 300 heterogeneous videos, produced by several European Broadcasters, to verify the effectiveness of the LG based logo detection and classification scheme. A total of 1800 TV frames extracted by the web scraper module were successfully processed. The video contents were made available by third party companies publishing them on web sharing platforms. The frames, containing different logos of various sizes, were processed comparing them with a 30 European TV Broadcasters logo dataset. No information about the exact localization of the logos within the video screen, was available. The results shows the effectiveness of the proposed solution. Given a candidate logo region, the framework has shows an average accuracy of 96% in logo recognition. Besides the implemented method appears quite robust to aspect ratio and scale

variations, coding artifacts and noise distortions.

We made a comparison of our approach by using a convolutional neural network (CNN). By referencing to the fig.3, the CNN approach covers points 3 to 7, the logo detection and recognition module. Due to the size limitation of our corpus, we decided to make usage of a transfer learning approach[14][15] using a convolutional neural network trained on the ImageNet database, to analyze the TV frames and extract the related deep features. On the basis on the data enrichment obtained we trained a logistic regression classifier, we tested various values of L1, L2 penalties and made a measure of the obtained accuracy of 61%, a value directly proportional to the dimension of the training dataset. We were not able within this research, to identify the break even point between the LG wavelet domain approach and the neural network, since the dataset had been already calculated. This comparison has however encouraged a reflection on the boundaries of neural networks usage in specific settings as the brand/logo recognition inside a video. Other studies have explored the theme of logo recognition via usage of neural network [16], but in the specificity of the boundaries of this specific task, the proposed framework offers better result in term of accuracy and it requires computing resources.

Conclusions

In this paper, we have investigated the challenging problem of TV logos extraction and recognition on the basis of the structural information associated with the Laguerre Gauss salient points. The proposed method is based on a nearest-neighbor matching scheme that labels the candidate logo region. The logo area under analysis is then compared to the logo data set, taking into account the similarity index of the LG descriptors associated with the key points, computed by local Fisher information analysis. Besides, the presented framework offers the inherent advantage to use the LG first order coefficients maps, that may be also used [17] to remove the detected logo by inpainting.

References

[1] Francis Gurry, Blue Sky Conference: Future Directions in Copyright Law

[2] Palfrey, John, et al. "Youth, creativity, and copyright in the digital age." (2009): 79-97.

[3] https://www.youtube.com/yt/copyright/

[4] J. Kleban, Xing Xie and Wei-Ying Ma, "Spatial pyramid mining for logo detection in natural scenes," 2008 IEEE International Conference on Multimedia and Expo, Hannover, 2008, pp. 1077-1080.

[5] Chang-Yu Lu, Myung-Cheol Roh, Seung-Yeon Kang, and Seong-Whan Lee. "Automatic logo transition detection in digital video contents". Pattern Anal. Appl. 15, 2 (May 2012), 175-187.

[6] H. Sahbi, L. Ballan, G. Serra and A. Del Bimbo, "Context-Dependent Logo Matching and Recognition," in IEEE Transactions on Image Processing, vol. 22, no. 3, pp. 1018-1031, March 2013.

[7] R. Boia, A. Bandrabur and C. Florea, "Local description using multi-scale complete rank transform for improved logo recognition," 2014 10th International Conference on Communications (COMM), Bucharest, 2014, pp. 1-4.

[8] SeleniumHq, web browser automation, http://www.seleniumhq.org

[9] Yan, Wei-Qi and Wang, Jun and Kankanhalli, Mohan S., Automatic video logo detection and removal, Multimedia Systems, v. 10,pp. 379-391, 2005.

[10] Licia Capodiferro, Elio D. Di Claudio, Giovanni Jacovitti, and Federica Mangiatordi. "Application of local Fisher information analysis to salient points extraction", Proc. of the Fifth IASTED International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA '08), 2008.

[11] A. Neri and G. Iacovitti. "Maximum likelihood localization of 2-d patterns in the gauss-laguerre transform domain: Theoretic framework and preliminary results". IEEE Transaction on Image Processing, 13,pp.7286, 2004.

[12] L. Capodiferro, F. Mangiatordi, E. Pallotti,Orientation and scale invariant salient points extraction,ICT 2008 Workshop on Semantic Multimodal Analysis of Digital Media, November 28, 2008,Lyon, France

[13] L. Capodiferro, E. D. Di Claudio, G. Jacovitti and A. Laurenti Local Orientation Estimation By Tomographic Hermite Slices The Fourth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications, SPPRA 2007

[14] Using Deep Learning for Image-Based Plant Disease Detection https://arxiv.org/abs/1604.03169

[15] Xue-Wen Chen, Xiaotong Lin, Big Data Deep Learning: Challenges and Perspectives Access, IEEE, Vol. 2 (2014), pp. 514-525 Yan, Wei-Qi and Wang, Jun and Kankanhalli, Mohan S., Automatic video logo detection and removal, Multimedia Systems, v. 10,pp. 379-391, 2005.

[16] Iandola, Forrest N., et al. "DeepLogo: Hitting logo recognition with the deep neural network hammer." arXiv preprint arXiv:1510.02131 (2015).

[17] Pallotti Emiliano, Capodiferro Licia, Mangiatordi Federica, Sit Paolo F., Smooth image inpainting by least square oriented edge prediction, Proc. SPIE 8295, Image Processing: Algorithms and Systems X; and Parallel Processing for Imaging Applications II, 82950H (2 February 2012).

Author Biography

Andrea Bernardini received his Dr. Ing. degree in Computer Engineering at the University of Rome "Roma Tre". In 2010, he was Visiting Researcher at the Institute for Computing, Information and Cognitive Systems (ICICS) of the University of British Columbia (UBC). In 2002 he joined The Fondazione Bordoni, where he works as researcher in Information processing and management Department. His research interests include User Experience, Data Mining and User Modeling.

Licia Capodiferro received her Dr. Ing. degree in Electronic Engineering from the University of Rome La Sapienza, Italy. In 1987 she joined the Fondazione Ugo Bordoni where she currently works as head of the Department of Information Processing and Management. Her main research interests are in the field of multimedia processing, with a focus on algorithms that allow the use of images and videos on the different types of terminals.

Federica Mangiatordi received the M.Sc. Degree in Electronic Engineering at University of Rome La Sapienza and the PhD in Electronic Materials, Optoelectronics and Microsystems from the University of Roma TRE. She works at Fondazione Ugo Bordoni from 2007. Her research interest concern multimedia retrieval, image restoration algorithms, novel metrics for full reference and no-reference image objective quality assessment.

Emiliano Pallotti received the Laurea Degree in Telecommunica-

tions Engineering at the University of Rome La Sapienza, Italy, and PhD in Electronic Materials, Optoelectronics and Microsystems from the University of Roma TRE. In 2007 he joined the Fondazione Ugo Bordoni where his research activities are in the field of on computational algorithms and video processing techniques based on multiresolution image representation in wavelet domain.