Multimedia Systems and Applications

Jenny Benois-Pineau
Patrick Le Callet *Editors*

Visual Content Indexing and Retrieval with Psycho-Visual Models

Springer

Multimedia Systems and Applications

Series editor
Borko Furht, Florida Atlantic University, Boca Raton, USA

More information about this series at http://www.springer.com/series/6298

Jenny Benois-Pineau • Patrick Le Callet
Editors

Visual Content Indexing and Retrieval with Psycho-Visual Models

Springer

Editors
Jenny Benois-Pineau
LaBRI UMR 5800, Univ. Bordeaux,
CNRS, Bordeaux INP
Univ. Bordeaux
Talence, France

Patrick Le Callet
LS2N, UMR CNRS 6004
Université de Nantes
Nantes Cedex 3, France

Multimedia Systems and Applications
ISBN 978-3-319-57686-2 ISBN 978-3-319-57687-9 (eBook)
DOI 10.1007/978-3-319-57687-9

Library of Congress Control Number: 2017946320

© Springer International Publishing AG 2017
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

We dedicate this book to our colleagues and friends, bio-physicists and neuroscientists. Aymar, Daniel, Mark... Thank you for our fruitful discussions on physiology and biological control of Human Visual System.

Preface

Since the early ages of Pattern Recognition, researchers try to make computers imitate the perception and understanding of visual content by humans. In the era of structural pattern recognition, the algorithms of contour and skeleton extrapolation in binary images tried to link missing parts using the principle of optic illusions described by Marr and Hildreth.

Modeling of Human Visual System (HVS) in perception of visual digital content has attracted a strong attention of research community in relation to the development of image and video coding standards, such as JPEG, JPEG2000, and MPEG1,2. The main question was how strongly and where in the image the information could be compressed without a noticeable degradation in the decoded content, thus ensuring quality of experience to the users. Nevertheless, the fundamental research on the borders of signal processing, computer vision, and psycho-physics continued and in 1998 has appeared the model of Itti, Koch and Niebur which has become the most popular model for prediction of visual attention. They were interested in both pixels-wise saliency and the scan-path, "static" and dynamic components. A tremendous amount of saliency models for still images and video has appeared during 2000 ties addressing both "low-level", bottom-up or stimuli-driven attention and high-level,"top-down", task-driven attention.

In parallel, content-based image an video indexing and retrieval community (CBIR and CVIR) has become strongly attached to the so-called "salient features", expressing signal singularities: corners, blobs, spatio-temporal jams in video. Using the description of the neighbourhood of these singularities, we tried to describe, retrieve and classify visual content addressing classical tasks of visual information understanding: similarity search in images, recognition of concepts, objects and actions. Since a few years these two streams have met. We are speaking today about "perceptual multimedia", "salient objects", and "interestingness" and try to incorporate this knowledge into our visual indexing and retrieval algorithms, we develop models of prediction of visual attention adapted to our particular indexing tasks... and we all use models of visual attention to drive recognition methods.

In this book we tried to give a complete state of the art in this highly populated and exploding research trend: visual information indexing and retrieval with psycho-visual models. We hope that the book will be interesting for researchers as well as PhD and master's students and will serve as a good guide in this field.

Bordeaux, France — Jenny Benois-Pineau
Nantes, France — Patrick Le Callet
March 2017

Acknowledgements

We thank the French National Research Network GDR CNRS ISIS for the support of scientific exchanges during our Workshops, and Souad Chaabouni and Boris Mansencal for their technical help in preparation of the manuscript of the book.

Contents

Contributors

Chokri Ben Amar REGIM-Lab LR11ES48, National Engineering School of Sfax, Sfax, Tunisia

Jenny Benois-Pineau LaBRI UMR 5800, Univ. Bordeaux, CNRS, Bordeaux INP, Univ. Bordeaux, Talence, France

Adrian G. Bors Department of Computer Science, University of York, York, UK

Lilian Calvet Simula Research Laboratory, Fornebu, Norvége

Axel Carlier IRIT, UMR 5505, Université Toulouse, Toulouse, France

Souad Chaabouni LaBRI UMR 5800, Univ. Bordeaux, CNRS, Bordeaux INP, Univ. Bordeaux, Talence, Cedex, France

Vincent Charvillat IRIT, UMR 5505, Université Toulouse, Toulouse, France

Frédéric Comby LIRMM, CNRS/Univ. Montpellier, Montpellier, France

Mihai Gabriel Constantin LAPI, University Politehnica of Bucharest, Bucharest, Romania

Vincent Courboulay L3i - University of La Rochelle, La Rochelle, France

Claire-Hélène Demarty Technicolor R&I, Rennes, France

Thanh-Toan Do Singapore University of Technology and Design, Singapore, Singapore

University of Science, Ho Chi Minh City, Vietnam

Michael Dorr Technical University Munich, Munich, Germany

Ngoc Q.K. Duong Technicolor R&I, Rennes, France

Christine Fernandez-Maloigne Xlim, University of Poitiers, CNRS, Poitiers, France

Syntyche Gbehounou Jules SAS, Blagnac, France

Pierre Gurdjos IRIT, UMR 5505, Université Toulouse, Toulouse, France

Bogdan Ionescu LAPI, University Politehnica of Bucharest, Bucharest, Romania

Pol Kennel IMFT, INPT/Univ. Toulouse, Toulouse, France

Patrick Le Callet LS2N, UMR CNRS 6004, Université de Nantes, Nantes Cedex 3, France

François Lecellier Xlim, University of Poitiers, CNRS, Poitiers, France

Karam Naser LS2N, UMR CNRS 6004, Polytech Nantes, University of Nantes, Nantes Cedex, France

Wei Tsang Ooi School of Computing, National University of Singapore, Singapore

Alex Papushoy Department of Computer Science, University of York, York, UK

William Puech LIRMM, CNRS/Univ. Montpellier, Montpellier, France

Arnaud Revel L3i - University of La Rochelle, La Rochelle, France

Vincent Ricordel LS2N, UMR CNRS 6004, Polytech Nantes, University of Nantes, Nantes Cedex, France

Mats Sjöberg Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland

Thierry Urruty Xlim, University of Poitiers, CNRS, Poitiers, France

Eleonora Vig German Aerospace Center, Oberpfaffenhofen, Germany

Hanli Wang Department of Computer Science and Technology, Tongji University, Shanghai, China

Akka Zemmari LaBRI UMR 5800, Univ. Bordeaux, CNRS, Bordeaux INP, Univ. Bordeaux, Talence, Cedex, France

Visual Content Indexing and Retrieval with Psycho-Visual Models

Patrick Le Callet and Jenny Benois-Pineau

Abstract The present chapter is an introduction to the book. The subject we propose has seen an exploded interest since last decade from research community in computer vision and multimedia indexing. From the field of video quality assessment where models of Human Visual System (HVS) were generally used to predict where humans will foveate and how will they perceive the degradation, these methods moved to classical Image and Video Indexing and retrieval tasks, recognition of objects, events, actions in images and video. In this book we try to give the most complete overview of the methods for visual information indexing and retrieval using prediction of visual attention or saliency. But also consider new approaches specifically designed for these tasks.

1 From Low to High Level Psycho Visual Models: Perceptual Computing and Applications

Along the last two decades, perceptual computing has emerged as a major topic for both signal processing and computer science communities. Taking care that many technologies produce signals for humans or process signals produced by humans, it is all the more important to consider perceptual aspects in the design loop. Whatever the uses cases, perceptual approaches rely on perceptual models that are supposed to predict and/or mimic some aspects of the perceptual system.

Such models are not trivial to obtain. Their development implies a multidisciplinary approach, in addition to signal processing of computer science encompassing neurosciences, psychology, physiology to name few. Perceptual modeling depends on the ability to identify the part of the system under study. In the case

P. Le Callet (✉)
LS2N UMR CNRS 6004, Université de Nantes, Nantes Cedex 3, France
e-mail: patrick.lecallet@univ-nantes.fr

J. Benois-Pineau
LaBRI UMR 5800, Univ. Bordeaux, CNRS, Bordeaux INP, Univ. Bordeaux, 351,
crs de la Liberation, F33405 Talence Cedex, France
e-mail: jenny.benois-pineau@u-bordeaux.fr

© Springer International Publishing AG 2017

J. Benois-Pineau, P. Le Callet (eds.), *Visual Content Indexing and Retrieval with Psycho-Visual Models*, Multimedia Systems and Applications,
DOI 10.1007/978-3-319-57687-9_1

of visual perception, sub-part of human visual system are easier to identify than some others, especially through psychophysics. With such approaches, relatively sufficient models have been successfully developed, mainly regarding "low level" of human vision. First order approximation for contrast perception such as Weber's law is a good and classic example, but we have been able to go much further, developing models for masking effects, color perception, receptive fields theory. In the late 1990s, there were already pretty advanced and practical perceptual models suitable for many image processing engineers. Most of them, such as Just Noticeable Difference (JND) models, are touching the visibility of signals and more specifically the visual differences between two signals. This knowledge is naturally very useful for applications such as Image quality prediction or image compression.

For years, these two applications have constituted a great playground for perceptual computing. They have probably pushed the evolution of perceptual models along the development of new immersive technologies (increasing resolution, dynamic range ...), leading not only to more advances JND models [19] but also to explore higher levels of visual perception.

Visual attention modeling is probably the best illustration of this trend, having concentrating massive efforts by both signal processing and computer science the last decade. From few papers in the mid 2000s, it is now a major topic covering several sessions in major conferences. High efforts on visual attention modeling can be legitimated also by applications angle. Knowing where humans are paying attention is very useful for perceptual tweaking of many algorithms: interactive streaming, ROI compression, gentle advertising [17]. Visual content indexing and retrieval field is not an exception and a lot of researchers have started to adopt visual attention modeling for their applications.

2 Defining and Clarifying Visual Attention

As the term *Visual Attention* has been used in a very wide sense, even more in the community that concerns this book, it requires few clarification. It is common to associate visual attention to eye gaze location. Nevertheless, eye gaze location do not necessarily fully reflect what human observers are paying attention to. One should first distinguish between overt and covert attention:

- **Overt attention** is usually associated with eye movements, mostly related to gaze fixation and saccades. It is easily observable nowadays with eye tracker devices, which record gaze tracking.
- **Covert attention**: William James [13] explained that we are able to focus attention to peripheral locations of interest without moving eyes. Covert attention is therefore independent of oculomotor commands. A good illustration is how a driver can remain fixating road while simultaneously covertly monitoring road signs and lights.

Even if overt attention and covert attention are not independent, over attention has been from far much more studied mostly because it can be measured in a straightforward way by using eye-tracking techniques. This is also one of the reasons why

the studies of computational modeling of visual attention are tremendously focused on overt attention. In that sense, visual attention is often seen in a simplified manner as a mechanism having at least the following basic components: (1) the selection of a region of interest in the visual field (2) the selection of feature dimensions and values of interest (3) the control of information flow through the network of neurons that constitutes the visual system; the shifting from one selected region to the next in time.

An important classification for Visual content indexing and retrieval field implies to distinguish between endogenous and exogenous mechanisms that drive visual attention. The **bottom-up** process is passive, reflexive, involuntary also known as exogenous as being driven by the signals, while the **top-down** process is active and voluntary and referred as endogenous attention. Attention can consequently either be task driven (Top-Down attention modeling) or feature driven (Bottom-Up attention modeling). The former is reflexive, signal driven, and independent of a particular task. It is driven involuntarily as a response to certain low-level features: motion, and in particular sudden temporal changes, is known to be dominant features in dynamic visual scenes whereas color and texture pop-outs represent the dominant features in the static scenes. Top-down attention, on the other hand, is driven by higher level cognitive factors and external influences, such as, semantic information, contextual effects, viewing task, and personal preference, expectations, experience and emotions. It is now widely known in the community that top-down effects are an inherent component of gaze behavior and these effects cannot be reduced or overcome even when no explicit task is assigned to the observers.

2.1 Interaction Between the Top-Down and Bottom-Up Attention Mechanisms

Itti et al. [12] describe the neurological backbone behind the top-down and bottom-up attention modeling as natural outcomes of the Inferotemporal cortex and Posterior parietal cortex based processing mechanisms respectively.

Whatever of the considered neurological model, it is more important in most usage of them, to appreciate the relative weights to be used or the mechanisms of interaction between these top-down and bottom-up approaches. Schill et al. [27] highlighted that humans gaze at regions where further disambiguation of information when required. After the gaze is deployed towards such a region, it is the bottom-up features which stand up by feature selection that helps achieve this goal. The work in [23] also highlights some important aspects of free-viewing in this regard, where the variation of the relative top-down versus bottom-up weight $\lambda(t)$ was examined as a function of time. While attention was initially found to be strongly bottom-up driven, there was a strong top-down affect in the range of 100–2000 ms. Later however the interaction between the two processes reach an equilibrium state.

2.2 The Concept of Perceived Importance/Interest: A Visual Attention Concept for Visual Content Indexing and Retrieval

From the application angle addressed in this book, it is desirable to get some models of visual attention. Despite their common goal of identifying the most relevant information in a visual scene, the type of relevance information that is predicted by visual attention models can be very different. While some of the models focus on the prediction of saliency driven attention locations, others aim at predicting regions-of-interest (ROI) at an object level.

Several processes are thought to be involved in making the decision for an ROI, including, attending and selecting a number of candidate visual locations, recognizing the identity and a number of properties of each candidate, and finally evaluating these against intentions and preferences, in order to judge whether or not an object or a region is interesting. Probably the most important difference between eye movement recordings and ROI selections is related to the cognitive functions they account for. It is very important to distinguish between three "attention" processes as defined by Engelke and Le Callet [6]:

- Bottom-up Attention: exogenous process, mainly based on signal driven visual attention, very fast, involuntary, task-independent.
- Top-down Attention: endogenous process, driven by higher cognitive factors (e.g. interest), slower, voluntary, task-dependent, mainly subconscious.
- Perceived Interest: strongly related to endogenous top-down attention but involving conscious decision making about interest in a scene.

Eye tracking data is strongly driven by both bottom-up and top-down attention, whereas ROI selections can be assumed to be mainly driven by top-down attention and especially perceived interest. It is the result of a conscious selection of the ROI given a particular task, providing the level of **perceived interest or perceptual importance**. Consequently, from a conceptual point of view, it might interesting to distinguish between two different types of perceptual relevance maps of a visual content: Importance versus Salience maps. While Salience refers to the pop-out effect of a certain feature: either temporally or spatially, importance maps indicates the perceived importance as it could be rated by human subjects. A saliency map is a probabilistic spatial signal, that indicates the relative probability with which the users regard a certain region. Importance maps on the other hand could be obtained by asking users to rate the importance of different objects in a scene.

2.3 Best Practices for Adopting Visual Attention Model

As stated before, the terms visual attention and saliency can be found in literature with various meaning. Whatever models adopted, researchers should be cautious and check if the selected model is designed to meet the requirements of the targeted

application. Moreover, one should also carefully verify the data on which models have been validated. In the context of Visual content indexing and retrieval applications, models touching concepts related to top down saliency, ROI and perceived interest/importance seem the more appealing. Nevertheless, while practically useful, it is very rare that these concepts are explicitly refereed as such, including some of the chapters in this book. The careful reader should be able to make this distinction when visual attention is concerned.

3 Use of Visual Attention Prediction in Indexing and Retrieval of Visual Content

Modeling the selective process of human perception of visual scenes represents an efficient way to drive the scene analysis towards particular areas considered 'of interest' or 'salient'. This is why it has become a very active trend in visual information indexing and retrieval [9]. Due to the use of saliency maps, the search for objects in images is more focused, thus improving the recognition performance and additionally reducing the computational burden. Even more, saliency methods can be naturally applied to all models which have been used up to now in these tasks, such as Bag-of-Visual-Words (BoVW) [25], sliding window approaches for visual object recognition [2, 31], image retrieval [4] or action recognition [32]. Saliency maps are used for generation of "object proposals" for recognition of objects in images and video with Deep Convolutional Neural Networks [5]. Hence in this book we give a large overview of the use of different visual attention models in fundamental tasks of visual information indexing: image and video querying and retrieval, action recognition, emotional analysis, visualization of image content. Models of visual attention, such as the one proposed by Itti et al. [12], Harel's graph implementation [10] are frequently used in literature for computing saliency maps. Nevertheless, as a function of target application and visual task, new forms of saliency can be predicted. Recently, the notion of saliency has been extended to the "interestingness" of visual content [24]. The latter can be understood globally for images and video fragments or locally, in which case it roughly delimits the area in image plane, where the objects of interest can be situated. This notion is also addressed in the present book.

We start with introduction of perceptual models in the problem of visual information retrieval at quite a general level. Visual textures represent areas in images which appears to be uniform from the perspective of human perception. It is difficult to speak here about salient areas, as this is the case in structural visual scenes with objects of interest. In chapter "Perceptual Texture Similarity for Machine Intelligence Applications" the authors are interested in how perceptual models can help in similarity matching of textures. The chapter reviews the theories of texture perception, and provides a survey about the up-to-date approaches for both static and dynamic textures similarity. The authors target video compression application.

In chapter "Deep Saliency: Prediction of Interestingness in Video with CNN" the authors propose a first approach to the prediction of areas-of-interest in video content. Deep Neural Networks have become winners in indexing of visual information. They have allowed achievement of better performances in the fundamental tasks of visual information indexing and retrieval such as image classification and object recognition. In fine-grain indexing tasks, namely object recognition in visual scenes, the CNNs have to evaluate multiple "object proposals", that is windows in the image plane of different size and location. Hence the problem of recognition is coupled with the problem of localization. In [8] a good analysis of recent approaches for object localization has been proposed, such as "regression approaches" as in [1, 28], and "sliding window approaches" as in [29] when the CNN processes multiple overlapping windows. The necessity to classify multiple windows makes the process of recognition heavy. The authors of Girshick et al. [8] proposed a so called Region-based convolutional network (R-CNN). They restrict number of windows using "selective search" approach [31] thus the classifier has to evaluate a limited number of (2K) "object proposals". Prediction of the interestingness of windows is another way to bound the search space. This prediction can be fulfilled with the same approach: a deep CNN trained on the ground truth of visual saliency maps build upon recorded gaze fixations of observers in a large-scale psycho-visual experiment.

In chapter "Introducing Image Saliency Information into Content Based Indexing and Emotional Impact Analysis" the authors are interested in the influence of pixel saliency in classical image indexing paradigms. They use the BoVW paradigm [22] which means building of image signature when selecting features in image plane, quantizing them with regard to a built dictionary and then computing the histogram of quantized features. The authors predict visual saliency of image pixels with Harel's model [10]. They compute a dense set of local image features by four methods: (1) Harris detector [11], (2) Harris-Laplace detector [18], (3) Difference-of-Gaussians (DOG) used in [16] to approximate Harris-Laplace detector and (4) Features from Accelerated Segment Test (FAST) detector [26]. They define "saliency" features on the basis of underlining saliency map. They experimentally show that when filtering out salient features, the drop of image retrieval accuracy is almost four times stronger compared to the removal of "non-salient" features. Such a study on a publicly available databases is a good experimental witness of the importance of saliency in selection of content descriptors and thus justifies the general trend.

Chapter "Saliency Prediction for Action Recognition" develops on the same idea. Here the problem of action recognition in video content is addressed. In order to reduce computational burden, the authors propose a non-uniform sampling of features accordingly to the saliency maps build on the gaze fixations available for a public Hollywood dataset. They follow the standard (improved) Dense Trajectories pipeline from [33–35]. Based on optical flow fields, trajectories are computed first, and then descriptors are extracted along these trajectories from densely sampled interest points. These descriptors comprise the shape of the trajectory, Histogram of Gradients (HOG), Histogram of Optical Flow (HOF), and Motion Boundary

Histograms (MBH). In order to exclude irrelevant trajectories corresponding to the background they compensate motion along the video sequence. Based on the a priori knowledge of video scenes they exclude detected humans from this compensation. Following the objective of selection of action-salient features, they compute several saliency maps. First of all, the central bias saliency map is computed. It expresses Buswell's central bias hypothesis that humans fixate the center of an image [3] or a video frame, and thus in video production the most important objects are situated in the center of video frames in footage. Then they compute an empirical saliency maps identifying smooth pursuit gaze fixation. These saliency maps are specifically relevant to the action recognition as humans perform smooth pursuit movement accommodating to the moving objects. Finally, an analytical saliency map using 2D +t Hessian is computed. Pruning of features is proposed considering Weibull distribution on saliency measures of computed maps. Their detailed studies on the Hollywood2 dataset convincingly show that using saliency—based pruning of features in a classical BoVW with Fisher encoding indexing scheme improves with regard to the base line when a smaller amount of descriptors is used.

In chapter "Querying Multiple Simultaneous Video Streams with 3D Interest Maps" the interestingness of an object in a visual scene is defined by the user. The method is designed for the selection of the best view of an object-of-interest in the visual scene in real-time when a 3D reconstruction of the scene is available. The user selects the region-of-interest on his/her mobile phone, then the 2D ROI is back-projected on a 3D view of the video scene which is obtained from independent cameras. The objects of interest are found inside a projection cone in a 3D scene and the view with the highest entropy is selected expressing the best contrasts in video. The framework is different from a classical Content-Based Image Retrieval schemes. It is designed for real-time and real-life scenarios where the quality of the video being captured in a querying process with the mobile phone can be very poor. Hence the intervention of the user is necessary do delimit the "saliency", which is region/object-of-interest in this case.

While in chapter "Querying Multiple Simultaneous Video Streams with 3D Interest Maps" the entropy is used for selection of the best view of the object-of-interest, in chapter "Information: Theoretical Model for Saliency Prediction—Application to Attentive CBIR" the authors propose an information—theoretical model of saliency itself. The novelty of the proposed work is to present an application of Frieden's well established information framework [7] that answers to the question: how to optimally extract salient information based on the low level characteristics that the human visual system provides? The authors integrate their biologically inspired approach into a real-time visual attention model and propose an evaluation which demonstrates the quality of the developed model.

Chapter "Image Retrieval Based on Query by Saliency Content" is devoted to the study on how the introduction of saliency in image querying could improve the results in terms of information retrieval metrics. They propose a Query by Salience Content Retrieval (QCSR) framework. The main parts of the QSCR system consist of image segmentation, feature extraction, saliency modelling and evaluating the distance in the feature space between a query image and a sample image from the

given pool of images [21]. The authors proposed to consider saliency of images at two levels: the local level is the saliency of segmented regions, the global level is the saliency defined by image edges. For querying image database they select salient regions using underlying Harel's (GBVS) saliency map [10]. To select salient regions to be used in a query the authors use the statistics which is a mean saliency value across a region. Salient regions are selected accordingly to the criterion of retrieval performance by thresholding of its histogram for the whole image partition. The authors use various thresholding methods including the well-known Otsu's method [20]. The querying is fulfilled by computation of Earth Mover Distance from regions of Query Image and the Database Image with saliency weighting. The global saliency expressed by the energy of contours is also incorporated into the querying process. They conduct multiple tests on CORELL 1000 and SIVAL databases and show that taking into account saliency allows for better top ranked results: more similar images are returned at the top of the rank list.

In chapter "Visual Saliency for the Visualization of Digital Paintings" the authors show how saliency maps can be used in a rather unusual application of visual content analysis, which is creation of video clips from art paintings for popularization of cultural heritage. They first built a saliency map completing Itti's model [12] with a saturation feature. Then the artist is selecting and weighting salient regions interactively. The regions of interest (ROIs) are then ordered accordingly to the central bias hypothesis. Finally, an oriented graph of salient regions is built. The graph edges express the order in which the regions will be visualized and the edges of the graph are weighted with transition times in the visualization process set by the artist manually. Several generated video clips were presented to eight naive users in a psycho-visual experiment with the task to score how the proposed video animation clip reflects the content of the original painting. The results, measured by the mean opinion score (MOS) metric, show that, in case of four-regions visualization, the MOS values for randomly generated animation clips and those generated with proposed method differ significantly up to 12%.

Finally, chapter "Predicting Interestingness of Visual Content" is devoted to the prediction of interestingness of multimedia content, such as image, video and audio. The authors consider visual interestingness from a psychological perspective. It is expressed by two structures "novelty-complexity" and a "coping potential". The former indicates the interest shown by subjects for new and complex events and the latter measures a subject's ability to discern the meaning of a certain event. From the content-driven, automatic perspective, the interestingness of content has been studied in a classical visual content indexing framework, selecting the most relevant image-based features within supervised learning (SVM) approach [30]. Interestingness of media content is a perceptual and highly semantic notion that remains very subjective and dependent on the user and the context. The authors address this notion for a target application of a VOD system, propose a benchmark dataset and explore the relevance of different features, coming from the most popular local features such as densely sampled SFIT to the latest CNN features extracted from fully connected layer fc7 and prob features from AlexNet Deep CNN [14]. The authors have conducted the evaluation of various methods

for media content interestingness assessment in the framework of the MediaEval Benchmarking Initiative for Media Evaluation [15]. In this evaluation campaign 12 groups were participating using prediction methods from SVM to Deep NNs with pre-trained data. The conclusion of the authors are that the task still remains difficult and open as the highest Mean Average Precision (MAP) metric values for image interestingness was 0.22 and for video interestingness it was only 0.18.

References

1. Agrawal, P., Girshick, B., Malik, J.: Analyzing the performance of multilayer neural networks for object recognition. In: Computer Vision - ECCV 2014–13th European Conference, Zurich, September 6–12 (2014), Proceedings, Part VII, pp. 329–344 (2014)
2. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2189–2202 (2012)
3. Buswell, G.T.: How People Look at Pictures. University of Chicago Press, Chicago, IL (1935)
4. de Carvalho Soares, R., da Silva, I.R., Guliato, D.: Spatial locality weighting of features using saliency map with a BoVW approach. In: International Conference on Tools with Artificial Intelligence, 2012, pp. 1070–1075 (2012)
5. de San Roman, P.P., Benois-Pineau, J., Domenger, J.-P., Paclet, F., Cataert, D., de Rugy, A.: Saliency driven object recognition in egocentric videos with deep CNN. CoRR, abs/1606.07256 (2016)
6. Engelke, U., Le Callet, P.: Perceived interest and overt visual attention in natural images. Signal Process. Image Commun. **39**(Part B), 386–404 (2015). Recent Advances in Vision Modeling for Image and Video Processing
7. Frieden, B.R.: Science from Fisher Information: A Unification, Cambridge edn. Cambridge University Press, Cambridge (2004)
8. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 142–158 (2016)
9. González-Díaz, I., Buso, V., Benois-Pineau, J.: Perceptual modeling in the problem of active object recognition in visual scenes. Pattern Recogn. **56**, 129–141 (2016)
10. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems, vol. 19, pp. 545–552. MIT, Cambridge (2007)
11. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proceedings of the 4th Alvey Vision Conference, pp. 147–151 (1988)
12. Itti, L., Koch, C.: Computational modelling of visual attention. Nat. Rev. Neurosci. **2**(3), 194–203 (2001)
13. James, W.: The Principles of Psychology. Read Books, Vancouver, BC (2013)
14. Jiang, Y.-G., Dai, Q., Mei, T., Rui, Y., Chang, S.-F.: Super fast event recognition in internet videos. IEEE Trans. Multimedia **177**(8), 1–13 (2015)
15. Larson, M., Soleymani, M., Gravier, G., Jones, G.J.F.: The benchmarking initiative for multimedia evaluation: MediaEval 2016. IEEE Multimedia **1**(8), 93–97 (2017)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**, 91–110 (2004)
17. Le Meur, O., Le Callet, P.: What we see is most likely to be what matters: visual attention and applications. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 3085–3088 (2009)
18. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: Proceedings of the 8th IEEE International Conference on Computer Vision, vol. 1, pp. 525–531 (2001)

19. Narwaria, M., Mantiuk, K.R., Da Silva, M.P., Le Callet, P.: HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images. J. Electron. Imaging **24**(1), 010501 (2015)
20. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
21. Papushoy, A., Bors, G.A.: Visual attention for content based image retrieval. In: 2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, 27–30 September 2015, pp. 971–975
22. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, Alaska, 24–26 June 2008
23. Rai, Y., Cheung, G., Le Callet, P.: Quantifying the relation between perceived interest and visual salience during free viewing using trellis based optimization. In: 2016 International Conference on Image, Video, and Multidimensional Signal Processing, vol. 9394, July 2016
24. Rayatdoost, S., Soleymani, M.: Ranking images and videos on visual interestingness by visual sentiment features. In: Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, 20–21 October 2016, CEUR-WS.org
25. Ren, X., Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video. In: IEEE Conference on Computer Vision and Pattern Recognition (2010)
26. Rosten, E., Drummond, T.: Fusing points and lines for high performance tracking. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2, pp. 1508–1511 (2005)
27. Schill, K., Umkehrer, E., Beinlich, S., Krieger, G., Zetzsche, C.: Scene analysis with saccadic eye movements: top-down and bottom-up modeling. J. Electron. Imaging **10**(1), 152–160 (2001)
28. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. CoRR, abs/1312.6229 (2013)
29. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, June 23–28, pp. 3626–3633 (2013)
30. Soleymani, M.: The quest for visual interest. In: ACM International Conference on Multimedia, New York, pp. 919–922 (2015)
31. Uijlings, J.R.R., Van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154–171 (2013)
32. Vig, E., Dorr, M., Cox, D.: Space-Variant Descriptor Sampling for Action Recognition Based on Saliency and Eye Movements, pp. 84–97. Springer, Firenze (2012)
33. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision (2013)
34. Wang, H., Kläser, A., Schmid, C., Liu, C.-L.: Action recognition by dense trajectories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176. IEEE, New York (2011)
35. Wang, H., Oneata, D., Verbeek, J., Schmid, C.: A robust and efficient video representation for action recognition. Int. J. Comput. Vis. 219–38 (2016)

Perceptual Texture Similarity for Machine Intelligence Applications

Karam Naser, Vincent Ricordel, and Patrick Le Callet

Abstract Textures are homogeneous visual phenomena commonly appearing in the visual scene. They are usually characterized by randomness with some stationarity. They have been well studied in different domains, such as neuroscience, vision science and computer vision, and showed an excellent performance in many applications for machine intelligence. This book chapter focuses on a special analysis task of textures for expressing texture similarity. This is quite a challenging task, because the similarity highly deviates from point-wise comparison. Texture similarity is key tool for many machine intelligence applications, such as recognition, classification, synthesis and etc. The chapter reviews the theories of texture perception, and provides a survey about the up-to-date approaches for both static and dynamic textures similarity. The chapter focuses also on the special application of texture similarity in image and video compression, providing the state of the art and prospects.

1 Introduction

Textures are fundamental part of the visual scene. They are random structures often characterized by homogeneous properties, such as color, orientation, regularity and etc. They can appear both as static or dynamic, where static textures are limited to spatial domain (like texture images shown in Fig. 1), while dynamic textures involve both the spatial and temporal domain Fig. 2.

Research on texture perception and analysis is known since quite a long time. There exist many approaches to model the human perception of textures, and also many tools to characterize texture. They have been used in several applications such

K. Naser (✉) • V. Ricordel
LS2N, UMR CNRS 6004, Polytech Nantes, University of Nantes, Rue Christian Pauc, BP 50609, 44306 Nantes Cedex 3, France
e-mail: karam.naser@univ-nantes.fr; vincent.ricordel@univ-nantes.fr

P. Le Callet
LS2N UMR CNRS 6004, Université de Nantes, Nantes Cedex 3, France
e-mail: patrick.lecallet@univ-nantes.fr

© Springer International Publishing AG 2017

J. Benois-Pineau, P. Le Callet (eds.), *Visual Content Indexing and Retrieval with Psycho-Visual Models*, Multimedia Systems and Applications,
DOI 10.1007/978-3-319-57687-9_2

Fig. 1 Example of texture images from VisTex Dataset

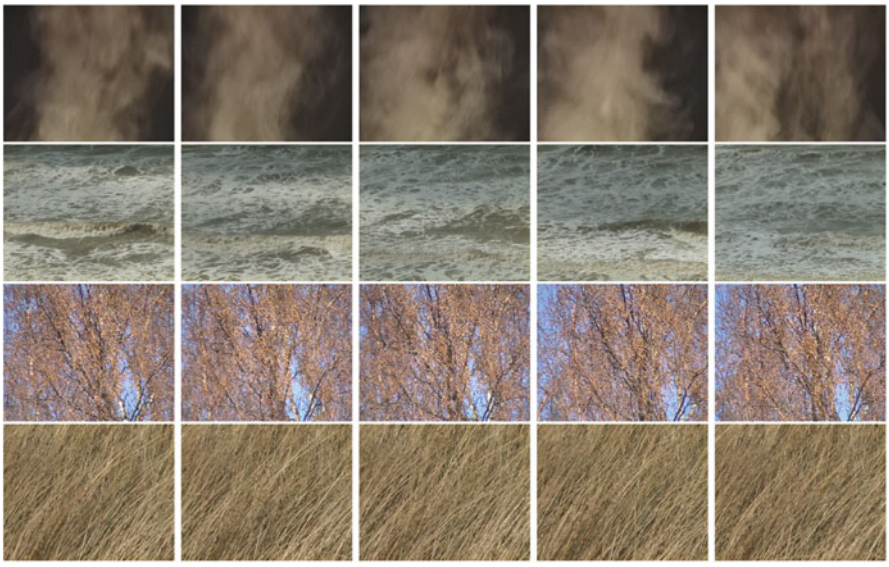

Fig. 2 Example of dynamic textures from DynTex Dataset [93]. *First row* represents the first frame, and *next rows* are frames after respectively 2 s

as scene analysis and understanding, multimedia content recognition and retrieval, saliency estimation and image/video compression systems.

There exists a large body of reviews on texture analysis and perception. For example, the review of Landy [57, 58] as well as the one from Rosenholtz [98] give a detailed overview of texture perception. Besides, the review of Tuceryan et al. in [117] covers most aspects of texture analysis for computer vision applications, such as material inspection, medical image analysis, texture synthesis and segmentation. On the other hand, the book Haindl et al. [45] gives an excellent review about modeling both static and dynamic textures. A long with this, there are also other reviews that cover certain scopes of texture analysis and perception, such as [29, 62, 88, 124, 135].

This chapter reviews an important aspect of texture analysis, which is texture similarity. This is because it is the fundamental tool for different machine intelligence applications. Unlike most of the other reviews, this covers both static and dynamic textures. A special focus is put on the use of texture similarity concept in data compression.

The rest of the chapter is organized as follows: Sect. 2 discusses about the meaning of texture in both technical and non-technical contexts. The details of texture perception, covering both static texture and motion perception, are given in Sect. 3. The models of texture similarity are reviewed in Sect. 4, with benchmarking tools in Sect. 5. The application of texture similarity models in image and video compression is discussed in Sect. 6, and the conclusion is given in Sect. 7.

2 What is Texture

Linguistically, the word texture significantly deviates from the technical meaning in computer vision and image processing. According to Oxford dictionary [86], the word refers to one of the followings:

1. *The way a surface, substance or piece of cloth feels when you touch it*
2. *The way food or drink tastes or feels in your mouth*
3. *The way that different parts of a piece of music or literature are combined to create a final impression*

However, technically, the visual texture has many other definitions, for example:

- *We may regard texture as what constitutes a macroscopic region. Its structure is simply attributed to pre-attentive patterns in which elements or primitives are arranged according to placement order* [110].
- *Texture refers to the arrangement of the basic constituents of a material. In a digital image, texture is depicted by spatial interrelationships between, and/or spatial arrangement of the image pixels* [2].
- *Texture is a property that is statistically defined. A uniformly textured region might be described as "predominantly vertically oriented", "predominantly small in scale", "wavy", "stubbly", "like wood grain" or "like water"* [58].
- *We regard image texture as a two-dimensional phenomenon characterized by two orthogonal properties: spatial structure (pattern) and contrast (the amount of local image structure)* [84].
- *Images of real objects often do not exhibit regions of uniform and smooth intensities, but variations of intensities with certain repeated structures or patterns, referred to as visual texture* [32].
- *Textures, in turn, are characterized by the fact that the local dependencies between pixels are location invariant. Hence the neighborhood system and the accompanying conditional probabilities do not differ (much) between various image loci, resulting in a stochastic pattern or texture* [11].
- *Texture images can be seen as a set of basic repetitive primitives characterized by their spatial homogeneity* [69].
- *Texture images are specially homogeneous and consist of repeated elements, often subject to some randomization in their location, size, color, orientation* [95].

- *Texture refers to class of imagery that can be characterized as a portion of infinite patterns consisting of statistically repeating elements* [56].
- *Textures are usually referred to as visual or tactile surfaces composed of repeating patterns, such as a fabric* [124].

The above definitions cover mostly the static textures, or spatial textures. However, the dynamic textures, unlike static ones, have no strict definition. The naming terminology changes a lot in the literature. The following names and definitions are summary of what's defined in research:

- Temporal Textures:
  1. They are class of image motions, common in scene of natural environment, that are characterized by structural or statistical self similarity [82].
  2. They are objects possessing characteristic motion with indeterminate spatial and temporal extent [97].
  3. They are textures evolving over time and their motion are characterized by temporal periodicity or regularity [13].
- Dynamic Textures:
  1. They are sequence of images of moving scene that exhibit certain stationarity properties in time [29, 104].
  2. Dynamic textures (DT) are video sequences of non-rigid dynamical objects that constantly change their shape and appearance over time[123].
  3. Dynamic texture is used with reference to image sequences of various natural processes that exhibit stochastic dynamics [21].
  4. Dynamic, or temporal, texture is a spatially repetitive, time-varying visual pattern that forms an image sequence with certain temporal stationarity [16].
  5. Dynamic textures are spatially and temporally repetitive patterns like trees waving in the wind, water flows, fire, smoke phenomena, rotational motions [30].
- Spacetime Textures:
  1. The term "spacetime texture" is taken to refer to patterns in visual spacetime that primarily are characterized by the aggregate dynamic properties of elements or local measurements accumulated over a region of spatiotemporal support, rather than in terms of the dynamics of individual constituents [22].
- Motion Texture:
  1. Motion textures designate video contents similar to those named temporal or dynamic textures. Mostly, they refer to dynamic video contents displayed by natural scene elements such as flowing rivers, wavy water, falling snow, rising bubbles, spurting fountains, expanding smoke, blowing foliage or grass, and swaying flame [19].

- Texture Movie:
  1. Texture movies are obtained by filming a static texture with a moving camera [119].
- Textured Motion:
  1. Rich stochastic motion patterns which are characterized by the movement of a large number of distinguishable or indistinguishable elements, such as falling snow, flock of birds, river waves, etc. [122].
- Video Texture:
  1. Video textures are defined as sequences of images that exhibit certain stationarity properties with regularity exhibiting in both time and space [42].

It is worth also mentioning that in the context of component based video coding, the textures are usually considered as details irrelevant regions, or more specifically, the region which is not noticed by the observers when it is synthesized [9, 108, 134].

As seen, there is no universal definition of the visual phenomena of textures, and there is a large dispute between static and dynamic textures. Thus, for this work, we consider the visual texture as:

A visual phenomenon, that covers both spatial and temporal texture, where spatial textures refer to homogeneous regions of the scene composed of small elements (texels) arranged in a certain order, they might exhibit simple motion such as translation, rotation and zooming. In the other hand, temporal textures are textures that evolve over time, allowing both motion and deformation, with certain stationarity in space and time.

3 Studies on Texture perception

3.1 Static Texture Perception

Static texture perception has attracted the attention of researchers since decades. There exists a bunch of research papers dealing with this issue. Most of the studies attempt to understand how two textures can be visually discriminated, in an effortless cognitive action known as pre-attentive texture segregation.

Julesz extensively studied this issue. In his initial work in [51, 53], he posed the question if the human visual system is able to discriminate textures, generated by a statistical model, based on the kth order statistics, and what is the minimum value of k that beyond which the pre-attentive discrimination is not possible any more. The order of statistics refers to the probability distribution of the of pixels values, in which the first order measures how often a pixel has certain color (or luminance value), while the second order measures the probability of obtaining a combination of two pixels (with a given distance) colors, and the same can be generalized for higher order statistics.