Smart Innovation, Systems and Technologies 76

Giuseppe De Pietro
Luigi Gallo
Robert J. Howlett
Lakhmi C. Jain *Editors*

Intelligent Interactive Multimedia Systems and Services 2017

KES International

Springer

Smart Innovation, Systems and Technologies

Volume 76

About this Series

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

More information about this series at http://www.springer.com/series/8767

Giuseppe De Pietro · Luigi Gallo
Robert J. Howlett · Lakhmi C. Jain
Editors

Intelligent Interactive Multimedia Systems and Services 2017

Springer

Editors
Giuseppe De Pietro
National Research Council of Italy (CNR-ICAR)
Institute for High-Performance Computing and Networking
Naples
Italy

Luigi Gallo
National Research Council of Italy (CNR-ICAR)
Institute for High-Performance Computing and Networking
Naples
Italy

Robert J. Howlett
Bournemouth University
Poole
UK

and

KES International
Shoreham-by-Sea
UK

Lakhmi C. Jain
University of Canberra
Canberra, ACT
Australia

and

Bournemouth University
Poole
UK

and

KES International
Shoreham-by-Sea
UK

ISSN 2190-3018 ISSN 2190-3026 (electronic)
Smart Innovation, Systems and Technologies
ISBN 978-3-319-59479-8 ISBN 978-3-319-59480-4 (eBook)
DOI 10.1007/978-3-319-59480-4

Library of Congress Control Number: 2017941493

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

Preface

Dear Readers,

We introduce to you a series of carefully selected papers presented during the 10th KES International Conference on Intelligent Interactive Multimedia Systems and Services (IIMSS-17).

At a time when computers are more widespread than ever, and computer users range from highly qualified scientists to non-computer expert professionals, intelligent interactive systems are becoming a necessity in modern computer systems. The solution of "one-fits-all" is no longer applicable to wide ranges of users of various backgrounds and needs. Therefore, one important goal of many intelligent interactive systems is dynamic personalization and adaptivity to users. Multimedia systems refer to the coordinated storage, processing, transmission, and retrieval of multiple forms of information, such as audio, image, video, animation, graphics, and text. The growth rate of multimedia services has become explosive, as technological progress matches consumer needs for content.

The conference took place as part of the Smart Digital Futures 2017 multi-theme conference, which groups AMSTA, IDT, InHorizons, InMed, SEEL with IIMSS in one venue. It was a forum for researchers and scientists to share work and experiences on intelligent interactive systems and multimedia systems and services. It included a general track and eight invited sessions.

The invited session "Processing visual data in intelligent systems: methods and applications" (Chaps. 1–8) specifically focuses on processing and understanding visual data in intelligent systems. The invited session "Cognitive Systems and Robotics" (Chaps. 9–20) focused on two main research areas, strictly related among them: adaptive and human-like cognitive systems, and artificial intelligence systems and cognitive robotics. The invited session "Big Data Management & Metadata" (Chaps. 21–24) focuses on models, techniques, and algorithms capable of dealing with the volume, velocity, variety, veracity, and value of big data. Differently, the invited session "Intelligent Big Data Analytics: Models, Techniques, Algorithms" (Chapter 25) discusses models, techniques, and algorithms for supporting intelligent analytics over big data in critical application contexts. The invited session

“Autonomous System” (Chaps. 26–29) considers technical and non-technical issues for what concerns intelligent, autonomous systems. The invited session “Mobile Data Analytics” (Chaps. 30–43) focuses on modeling, processing, and analyzing data generated by mobile devices, positioning technologies, and mobile users’ activities. The invited session “Smart Environments and Information Systems” (Chaps. 44–49) provides insight into the most recent efforts in the field of information systems operating in dynamic environments. The invited session “Innovative Information Services for Advanced Knowledge Activity” (Chaps. 50–53) focuses on novel functionalities for information services. Finally, the general track (Chaps. 54–57) focuses on topics related to image processing algorithms and image processing-based rehabilitation and recommender systems.

Our gratitude goes to many people who have greatly contributed to putting together a fine scientific program and exciting social events for IIMSS 2017. We acknowledge the commitment and hard work of the program chairs and the invited session organizers. They have kept the scientific program in focus and made the discussions interesting and valuable. We recognize the excellent job done by the program committee members and the extra reviewers. They evaluated all the papers on a very tight schedule. We are grateful for their dedication and contributions. We could not have done it without them. More importantly, we thank the authors for submitting and trusting their work to the IIMSS conference.

We hope that readers will find in this book an interesting source of knowledge in fundamental and applied facets of intelligent interactive multimedia and, maybe, even some motivation for further research.

The editors

Organization

Honorary Chairs

Toyohide Watanabe	Nagoya University, Japan
Lakhmi C. Jain	University of Canberra, Australia and Bournemouth University, UK

Co-General Chairs

Giuseppe De Pietro	National Research Council of Italy, Italy
Luigi Gallo	National Research Council of Italy, Italy

Executive Chair

Robert J. Howlett	University of Bournemouth, UK

Programme Chair

Antonino Mazzeo	University of Naples Federico II, Italy

Publicity Chair

Giuseppe Caggianese	National Research Council of Italy, Italy

Invited Session Chairs

Processing Visual Data in Intelligent Systems: Methods and Applications

Francesco Bianconi	Università degli Studi di Perugia, Italy
Elena González	Universidade de Vigo, Spain
Manuel Ángel Aguilar	Universidad de Almería, Spain

Cognitive Systems and Robotics

Ignazio Infantino	National Research Council of Italy, Italy
Massimo Esposito	National Research Council of Italy, Italy

Big Data Management and Metadata

Flora Amato	University of Naples Federico II, Italy
Vincenzo Moscato	University of Naples Federico II, Italy

Intelligent Big Data Analytics: Models, Techniques, Algorithms

Alfredo Cuzzocrea	University of Trieste, and ICAR-CNR, Italy

Autonomous System

Milan Simic	RMIT University, Australia

Mobile Data Analytics

Jalel Akaichi	University of Tunis, Tunisia, and King Khalid University, Saudi Arabia

Smart Environments and Information Systems

Rafael H. Bordini	FACIN-PUCRS, Brazil
Massimo Cossentino	National Research Council of Italy, Italy
Marie-Pierre Gleizes	University Paul Sabatier of Toulouse, France
Luca Sabatucci	National Research Council of Italy, Italy

Innovative Information Services for Advanced Knowledge Activity

Koichi Asakura	Daido University, Japan
Toyohide Watanabe	Nagoya Industrial Research Institute, Japan

International Programme Committee

Manuel Ángel Aguilar	Universidad de Almería, Spain
Flora Amato	Università degli Studi di Napoli Federico II, Italy
Marco Anisetti	Università degli Studi di Milano, Italy
Koichi Asakura	Daido University, Japan
Jalel Akaichi	University of Tunis, Tunisia, and King Khalid University, Saudi Arabia
Vivek Bannore	KES UniSA, Australia
V. Bellandi	Università degli Studi di Milano, Italy
Monica Bianchini	University of Perugia, Italy
Rafael H. Bordini	FACIN-PUCRS, Brazil
Helder Coelho	Mind-Brain College, BioISI, University of Lisbon, Portugal
Luigi Coppolino	Università degli Studi di Napoli “Parthenope”, Italy
Massimo Cossentino	National Research Council of Italy, Italy
Giovanni Cozzolino	Università degli Studi di Napoli Federico II, Italy
Alfredo Cuzzocrea	University of Trieste and ICAR-CNR, Italy
Salvatore D’Antonio	Università degli Studi di Napoli “Parthenope”, Italy
Ernesto Damiani	Università degli Studi di Milano, Italy
Mario Doeller	University of Applied Science Kufstein Tirol, Austria
Dinu Dragan	University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
Massimo Esposito	National Research Council of Italy, Italy
Margarita Favorskaya	Siberian State Aerospace University, Russia
Marie-Pierre Gleizes	University Paul Sabatier of Toulouse, France
Christos Grecos	Central Washington University, USA
Elena González	Universidade de Vigo, Spain
Vincent Hilaire	Université de Belfort-Montbeliard, France
Katsuhiro Honda	Osaka Prefecture University, Japan
Hsiang-Cheh Huang	National University of Kaohsiung, Taiwan
Ignazio Infantino	National Research Council of Italy, Italy
Gwanggil Jeon	Xidian University, China
Dimitris Kanellopoulos	Department of Mathematics, University of Patras, Greece
Chengjun Liu	New Jersey Institute of Technology, USA
Marian Cristian Mihaescu	University of Craiova, Romania

Lyudmila Mihaylova	University of Sheffield, UK
Vincenzo Moscato	Università degli Studi di Napoli Federico II, Italy
Francesco Moscato	Università degli Studi della Campania, Italy
Vincent Oria	New Jersey Institute of Technology, USA
Radu-Emil Precup	Politehnica University of Timisoara, Romania
Antonio Maria Rinaldi	Università degli Studi di Napoli Federico II, Italy
Luigi Romano	Università degli Studi di Napoli “Parthenope”, Italy
Luca Sabatucci	National Research Council of Italy, Italy
Mohammed Sadgal	Cadi Ayyad University, Morocco
Milan Simic	RMIT University, School of Engineering, Australia
Mariacarla Staffa	Università degli Studi di Napoli “Parthenope”, Italy
Claudio Sterle	Università degli Studi di Napoli Federico II, Italy
Porfirio Tramontana	Università degli Studi di Napoli Federico II, Italy
Taketoshi Ushiama	Kyushu University, Japan
Rosa Vicari	Universidade Federal do Rio Grande do Sul, Brazil
Toyohide Watanabe	Nagoya Industrial Science Research Institute, Japan
Alicja Wieczorkowska	Polish-Japanese Academy of Information Technology, Poland

Contents

Hand-Designed Local Image Descriptors vs. Off-the-Shelf CNN-Based Features for Texture Classification: An Experimental Comparison

Raquel Bello-Cerezo[1(✉)], Francesco Bianconi[1], Silvia Cascianelli[1], Mario Luca Fravolini[1], Francesco di Maria[1], and Fabrizio Smeraldi[2]

[1] Department of Engineering, Università degli Studi di Perugia, Via G. Duranti 93, 06135 Perugia, PG, Italy
{raquel.bellocerezo,silvia.cascianelli}@studenti.unipg.it, bianco@ieee.org, {mario.fravolini,francesco.dimaria}@unipg.it

[2] School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, UK
f.smeraldi@qmul.ac.uk

Abstract. Convolutional Neural Networks have proved extremely successful in object classification applications; however, their suitability for texture analysis largely remains to be established. We investigate the use of pre-trained CNNs as texture descriptors by tapping the output of the last fully connected layer, an approach that has proved its effectiveness in other domains. Comparison with classical descriptors based on signal processing or statistics over a range of standard databases suggests that CNNs may be more effective where the intra-class variability is large. Conversely, classical approaches may be preferable where classes are well defined and homogeneous.

Keywords: Convolutional Neural Networks · Image classification · Texture · Local Binary Patterns

1 Introduction

Texture, along with colour, shape and gloss, is a fundamental visual feature of objects, materials and scenes. As a consequence, texture analysis plays an important role in several computer vision applications, such as image classification, content-based image retrieval, medical image analysis, surface inspection and remote sensing. Research on texture has been intense for more than forty years now: ideally, we could trace its origin as far back as 1973, when Haralick's seminal work on co-occurrence matrices [12] was first published. Since then a lot of different textures descriptors have been proposed in the literature: so many that Xie and Mirmehdi referred to them as 'a galaxy' [29]. Among them, methods based on signal processing like Gabor filters and wavelets dominated the scene

G. De Pietro et al. (eds.), *Intelligent Interactive Multimedia Systems and Services 2017*, Smart Innovation, Systems and Technologies 76, DOI 10.1007/978-3-319-59480-4_1

for a while, whereas in the last two decades statistical and rank-based features have become more popular. The bag-of-features paradigm [30] has also become the prominent aggregation strategy.

In recent years the appearance of Convolutional Neural Networks (CNNs) [14] represented a major breakthrough that changed the outlook for the pattern recognition field. This new paradigm for image analysis proved to consistently outperform pre-existing methods in a number of applications including object image recognition and scene classification [14,26]. Central to this scheme is the ability to learn complex image-to-object or image-to-feature mappings starting from very large datasets of labelled images. More importantly, pre-trained CNNs have also showed to be able to generalise quite well to datasets different from those they are trained on [8,26,31], a feature that makes them amenable to being used 'out of the box' in a potentially large number of applications. Yet the real effectiveness of CNN-based methods with *fine-grained* images – such as texture – is still subject of debate. Most of the related literature, that we briefly review in Sect. 2, is in fact rather new, and the results are far from being consolidated.

In this work we investigate the effectiveness of CNNs compared with classic local image descriptors such as Local Binary Patterns and variants, Gabor filters and grey-level co-occurrence matrices for texture classification. Specifically, we are interested in determining the potential of pre-trained CNNs when used as feature extractors in an off-the-shelf manner, relying directly on the pooling effect of the fully connected layers of the network. This avoids the added complexity of the separate pooling stages appearing in some related studies, that we review in Sect. 2). In the remainder of the paper we describe the materials (Sect. 3) and methods (Sect. 4) used in this study. We discuss the experimental set-up and the results in Sect. 5 and conclude the paper with some final considerations (Sect. 6) and directions for future studies (Sect. 7).

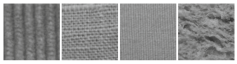
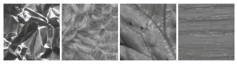
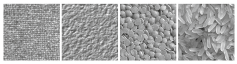
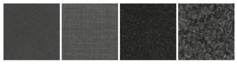
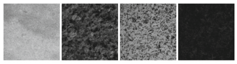
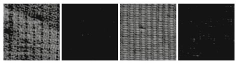
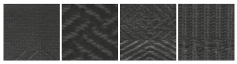
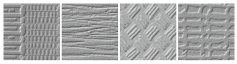
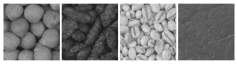
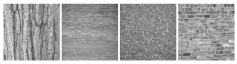
2 Related Research

Convolutional Neural Networks have been attracting increasing research interest in the computer vision community: suffice it to say that Krizhevsky *et al.*'s milestone work [14] has been so far cited more than 2700 times[1] since its publication in 2012.

In the field of texture analysis CNN-based methods have been receiving increasing attention. Cimpoi *et al.* [8] is the first in-depth investigation of the transferability of CNN models to the texture domain. The proposed solution (FV-CNN), however, entails complex and time-consuming pre- and post-processing procedures (respectively repeated image rescaling and Fisher vector pooling) that actually make it a new texture descriptor on its own rather than a direct application of CNNs to textures. A potential drawback of this solution is also the huge number of features produced (65K) which may represent a limit in many practical applications. Andrearczy and Whelan [1] recently improved on this idea and proposed a pooling scheme which relies on a lower number of features.

[1] Source: Scopus®; visited on Januray 18, 2017.

Table 1. Round-up table of the datasets used in the experiments.

ID	Name	No. of classes	No. of samples per class	Sample images
1	KTH-TIPS	10	81	
2	KTH-TIPS2b	11	432	
3	Kylberg	28	160	
4	Kylberg-Sintorn	25	6	
5	MondialMarmi	25	16	
6	Outex-00013	68	20	
7	Outex-00014	68	60	
8	PerTex	334	16	
9	RawFooT	68	184	
10	UIUC	25	40	

An interesting comparison between LBP variants and CNN-based features – though once again obtained by vector pooling – was recently presented by Liu *et al.* [19]. Here the authors find that the best performance is obtained by an LBP variant known as Median Robust Extended Local Binary Patterns (MRELBP). Of late, an experimental evaluation of colour texture descriptors under variable lighting conditions – including CNN-based features – was proposed by Cusano *et al.* [11]. Their approach consists of generating a texture descriptor by using, as image features, the output of the last fully-connected layer of a CNN. The main advantage of this strategy is that it generates significantly fewer features than the pooling method, and can be considered the model that best fits the idea of off-the-shelf use of CNNs for texture analysis. Finally, it is worth noting that in later experiments the same Cusano *et al.* [10] found that Fisher vector pooling produced worse results than were obtained by directly using CNNs features, probably due to the high number of features generated by FV-CNN.

3 Materials

We considered 10 datasets of texture images: (1) KTH-TIPS; (2) KTH-TIPS2b; (3) Kylberg Texture Dataset; (4) Kylberg-Sintorn Rotation Dataset; (5) MondialMarmi; (6) Outex-00013; (7) Outex-00014; (8) Pertex; (9) RawFooT and (10) UIUC. The main features of each dataset are detailed in Sect. 3.1 and summarised in Table 1.

3.1 Datasets

KTH-TIPS [13,15] features 10 classes of materials: aluminum foil, bread, corduroy, cotton, cracker, linen, orange peel, sandpaper, sponge and styrofoam. Images of each material were taken under different viewpoints and illumination conditions, giving 81 images for each class.

KTH-TIPS2b [6,15] is an extension of KTH-TIPS and contains 11 types of materials: aluminum foil, brown bread, corduroy, cork, cotton, cracker, lettuce, linen, white bread, wood and wool. Four samples for each class were acquired under varying scale, illumination and pose resulting in 432 images for each class.

Kylberg Texture Dataset (v. 1.0) [16] contains 28 texture classes such as fabric, natural stone, grains and seeds. There are 160 images for each class; the samples contain no variation in scale, rotation or illumination.

Kylberg-Sintorn Rotation Dataset [17,18] is a collection of 25 classes of heterogeneous materials including food (seeds and sugar), textiles (wool and knitwear) and tiles, with one image per class. The image samples used in our experiments contain no variation in scale, rotation or illumination.

MondialMarmi (v 2.0) [2,21] is a visual catalogue of polished natural stone products (marble and granites) featuring 25 classes of commercial denominations (e.g., *Azul Platino*, *Bianco Sardo*, *Rosa Porriño*, etc.) with four samples per class – each sample representing one tile. The images were acquired at fixed scale, in controlled illumination conditions and under different rotation angles. In our experiments we only used non-rotated images and subdivided each of them into four non-overlapping sub-images, thus obtaining 16 samples per class.

Outex-00013 contains the same 68 texture classes as Outex's test suite TC-00013, i.e. a collection of heterogeneous materials such as grains, fabric, natural stone and wood (see also Ref. [22] for details). There are 20 image samples for each class which were acquired at fixed scale, rotation angle and under invariable illumination conditions.

Outex-00014 is composed of the same classes as Outex-00013; in this case, however, each sample was acquired under three different lighting sources. As a consequence there are 60 samples for each class instead of 20. Note that in order to maintain the same evaluation protocol for all the datasets (see Sect. 5) the splits used in our experiments are different from those provided respectively with the TC-00013 and TC-00014 test suites.

PerTex [9,24] includes 334 texture classes representing heterogeneous materials such as embossed vinyl, woven wall coverings, carpet, rugs, fabric, building materials, product packaging, etc. The images were obtained by calculating the height-maps of the samples first, then by relighting them in order to remove variations due to reflectance. The results is a dataset of highly homogeneous textures – some of which are very similar to each other, a feature that makes this a very challenging dataset.

Table 2. Summary table of the image descriptors considered in the experiments

Method	Acronym	No. of features
Hand-designed local image descriptors		
Completed Local Binary Patterns	CLPB	324
Gradient-based Local Binary Patterns	GLBP	108
Improved Local Binary Patterns	ILBP	213
Local Binary Patterns	LBP	108
Local Ternary Patterns	LTP	216
Texture Spectrum	TS	2502
Gabor Filters	$\text{Gabor}_{4,6}^{rw}$	48
	$\text{Gabor}_{4,6}^{cn}$	48
	$\text{Gabor}_{5,7}^{rw}$	70
	$\text{Gabor}_{5,7}^{cn}$	70
Grey-level co-occurrence matrices	GLCM	60
CNN-based features		
CNN-imagenet-caffe-alex	Caffe-AlexNet	4096
CNN-imagenet-vgg-fast	VGG-F	4096
CNN-imagenet-vgg-medium	VGG-M	4096
CNN-imagenet-vgg-slow	VGG-S	4096
CNN-imagenet-vgg-medium-128	VGG-M-128	128
CNN-imagenet-vgg-medium-1024	VGG-M-1024	1024
CNN-imagenet-vgg-medium-2048	VGG-M-2028	2048
CNN-imagenet-vgg-verydeep-16	VGG-VD-16	4096
CNN-vgg-face	VGG-Face	4096

RawFooT [11,25] contains 68 classes of different types of food such as grain, fish, fruit and meat. There are 46 image samples for each class, each sample having been acquired under 46 different lighting conditions, whereas scale and rotation angle are invariable. In our experiments we subdivided each sample into four non-overlapping images of smaller size, therefore obtaining $46 \times 4 = 184$ samples for each class.

UIUC features 25 classes of heterogeneous materials and objects such as bark, wood, water, granite, marble, floor, pebbles, wall, brick, glass, carpet, upholstery, wallpaper, fur, knit, corduroy and plaid. There are 40 samples for each class, and within each class there is a lot of variability due to significant changes in the imaging conditions (i.e. rotation, scale and viewpoint) and warped surfaces.

4 Methods

We included in the experiments 11 hand-designed local image descriptors – specifically: six variants of Local Binary Patterns, four sets of features from

Gabor filters and one from grey-level co-occurrence matrices. On the network side we had nine sets of CNN-based features from as many pre-trained CNNs. The comparison was carried out on grey-scale images, therefore discarding colour information altogether. Details about settings and implementation are provided in the following subsections. Table 2 summarises the whole set of image descriptors and lists the number of features generated by each method.

4.1 Hand-Designed Local Image Descriptors

We took into account the following LBP variants: Completed Local Binary Patterns, Gradient-based Local Binary Patterns, Improved Local Binary Patterns, Local Binary Patterns, Local Ternary Patterns and Texture Spectrum (please refer to Ref. [5] for details). For each descriptor we concatenated the rotation-invariant features (e.g. LBP^{ri}) computed over three concentric, non-interpolated, eight-pixel circles respectively of radius 1px, 2px and 3px.

Gabor features [3] were computed using two filter banks: one with four frequencies and six orientations, and the other with five frequencies and seven orientations, which in the remainder we respectively indicate as $\text{Gabor}_{4,6}$ and $\text{Gabor}_{5,7}$. In both cases we set the maximum frequency to 0.5px^{-1}, the frequency spacing to half-octave, the spatial frequency bandwidth and the aspect ratio to 0.5. We considered both raw and contrast-normalised filter output (in the latter case the filter responses for one point in all frequencies and rotations were normalized to sum one). In the reminder we indicate the two versions respectively with superscripts 'rw' and 'cn'. Image features were in all cases the mean and standard deviation of the magnitude of the Gabor-transformed images.

For the co-occurrence features we used 12 displacement vectors resulting from combining three distances (i.e. 1px, 2px and 3x – just as for LBP variants) and four standard orientations (i.e. 0°, 45°, 90° and 135°). From each matrix we extracted the following global statistics as image features: *contrast*, *correlation*, *energy*, *entropy* and *homogeneity* (see also Ref. [4] for details).

4.2 CNN-Based Features

CNN-based features were computed using nine pre-trained Convolutional Neural Networks. The image processing pipeline included a pre-processing step whereby the input images were converted to grey-scale first, then resized through bicubic interpolation to fit the input dimension of each network – which for all the networks considered here was $224\text{px} \times 224\text{px}$. The nets were fed by dealing the resized, grey-scale images to the three colour input channels. Following the approach proposed by Cusano *et al.* [11] we used as texture features the L_2-normalised output of the last fully-connected layer. The implementation was based on the MatConvNet platform [20, 28]. The main features of each network are summarised here below.

- **Caffe-AlexNet**: a MatConvNet porting of AlexNet, the architecture originally proposed by Krizhevsky *et al.* [14]. It is composed of eight layers, of which the first five are convolutional and the remaining three fully-connected.

Table 3. Overall accuracy by descriptor and dataset. Boldface figures indicate the best result for each dataset. Dataset IDs are listed in Table 1.

Descriptor	Dataset ID									
	1	2	3	4	5	6	7	8	9	10
Hand-designed local image descriptors										
CLBP	90.5	93.0	99.2	92.5	97.5	77.7	79.9	96.5	90.0	76.5
GBLBP	86.9	89.3	98.2	95.4	97.2	81.9	82.9	95.7	88.1	60.0
ILBP	89.9	91.7	99.1	95.8	**97.7**	**83.6**	**85.7**	96.5	93.1	72.0
LBP	87.7	89.7	98.0	90.1	96.7	78.2	80.5	95.5	90.1	60.5
LTP	87.8	89.8	98.1	90.1	96.7	79.0	81.7	95.6	90.5	60.6
TS	85.7	91.4	98.6	91.8	96.4	77.8	80.5	**97.3**	91.6	67.9
$\text{Gabor}^{rw}_{4,6}$	75.9	82.7	94.0	85.4	87.9	64.3	67.5	90.9	72.3	51.1
$\text{Gabor}^{rw}_{5,7}$	77.8	84.9	96.3	88.7	89.7	66.8	70.0	92.4	74.0	53.5
$\text{Gabor}^{cn}_{4,6}$	75.1	78.6	93.9	83.3	82.8	70.3	77.1	91.8	88.6	40.4
$\text{Gabor}^{rw}_{5,7}$	75.6	79.9	96.2	86.0	88.0	71.7	78.8	92.7	92.2	42.2
GLCM	75.4	80.5	97.2	92.9	89.9	65.3	68.2	92.9	74.4	52.8
CNN-based features										
Caffe-AlexNet	94.4	96.5	97.8	99.6	91.7	79.9	84.2	89.0	96.7	82.9
VGG-M	95.1	97.0	99.5	98.6	92.1	80.6	84.9	94.2	97.9	89.8
VGG-F	93.3	96.0	98.8	**99.7**	91.4	80.1	84.2	91.1	97.3	86.5
VGG-S	94.5	97.3	**99.7**	98.7	92.8	79.4	84.7	93.3	97.8	91.0
VGG-M-128	90.5	93.2	98.1	96.9	85.2	76.6	81.7	86.6	97.0	81.3
VGG-M-1024	94.4	96.6	99.4	99.2	91.0	79.2	84.3	93.3	97.7	88.3
VGG-M-2048	94.4	96.8	99.4	98.7	92.4	79.6	84.4	94.0	97.8	89.1
VGG-VD-16	**96.8**	**97.8**	99.5	99.5	93.8	80.6	85.6	93.4	**98.3**	**93.3**
VGG-face	86.5	87.1	92.1	97.5	85.0	71.4	82.0	68.7	95.5	57.7

- **VGG-F**, **VGG-M** and **VGG-S**: three networks all consisting of five convolutional and three fully-connected layers. The main differences are the size of the filters, the stride and the dimension of the pooling windows (‘F’, ‘M’ and ‘S’ respectively stand for *fast*, *medium* and *slow* – see also Ref. [7] for details).
- **VGG-M-128**, **VGG-M-1024** and **VGG-M-2048**: three variations of VGG-M with a lower-dimensional last fully-connected layer [7].
- **VGG-VD-16**: a deep network featuring 13 convolutional and three fully-connected layers [27].
- **VGG-Face**: a network designed for face recognition composed of eight convolutional and three fully-connected layers [23].

Apart from VGG-Face, which understandably was trained on faces [23], all the other networks were trained for object recognition.

5 Experiments and Results

To comparatively assess the effectiveness of the image descriptors presented in Sect. 4 we ran a supervised image classification experiment using the 1-NN classifier with L_1 distance. Accuracy estimation was based on split-sample validation with stratified sampling where 1/4 of the samples of each class was used to train the classifier and the remaining 3/4 to test it. The estimated accuracy was the ratio between the number of samples correctly classified and the total number of samples of the test set. For a stable estimation of the classification error we averaged the results (see Table 3) over 100 random splits.

The results are interesting and show a trend strongly dependent on the dataset used. In six datasets out of 10, CNN-based features outperformed the hand-designed methods (though in dataset #3 the margin is narrow); whereas the reverse occurred in four datasets out of 10 (though again by a narrow margin in dataset #7). CNN-based features seemed to be more effective when there was high intra-class variability due to changes in viewpoint/scale/appearance: paradigmatic and impressive is the 93.3% attained by VGG-VD-16 on dataset UIUC – a notoriously difficult one. By contrast, hand-designed image descriptors appeared to be more comfortable with homogeneous, fine-grained textures and little intra-class variability – as for instance in datasets #5 and #8. Within this group of methods, LBP variants clearly outperformed Gabor filters and GLCM.

6 Conclusions

Convolutional Neural Networks represented a major breakthrough in computer vision, having significantly improved the state of the art in many applications. Originally developed for object and scene classification, the approach proved effective in other domains as well, for example face recognition. It is however still a subject of debate whether this paradigm is amenable to being successfully applied to fine-grained images – i.e. texture. In this work we have carried out a comparison between some classic local image descriptors and off-the-shelf CNN-based features from an array of pre-trained nets. Our results were split, showing that though CNN-based features performed generally well, they were in some cases outperformed by state-of-the-art hand-designed descriptors. More specifically, our findings seem to suggest that CNNs performed better when there was high intra-class variability, whereas LBP variants provided better results with homogeneous, fine-grained textures with low intra-class variability.

7 Limitations and Future Work

The results presented here are promising and should be validated in a broader cohort of experiments. Importantly, our investigation was limited to grey-scale images, therefore the contribution of colour to image classification wasn't considered. Likewise, disturbing effects such as rotation and noise were not investigated. Assessing the effectiveness of more complex pooling schemes for CNN-based features (e.g. Fisher vectors) is also another important question for future studies.

Acknowledgements. This work was partially supported by the Department of Engineering at the Università degli Studi di Perugia, Italy, under project *BioMeTron* – Fundamental research grant D.D. 20/2015 and by the Spanish Government under project AGL2014-56017-R.

References

1. Andrearczyk, V., Whelan, P.F.: Using filter banks in convolutional neural networks for texture classification. Pattern Recogn. Lett. **84**, 63–69 (2016)
2. Bianconi, F., Bello, R., Fernández, A., González, E.: On comparing colour spaces from a performance perspective: application to automated classification of polished natural stones. In: Murino, V., Puppo, E., Sona, D., Cristani, M., Sansone, C. (eds.) New Trends in Image Analysis and Processing, ICIAP 2015 Workshops, Genoa, Italy. LNCS, vol. 9281, pp. 71–78. Springer (2015)
3. Bianconi, F., Fernández, A.: Evaluation of the effects of Gabor filter parameters on texture classification. Pattern Recogn. **40**(12), 3325–3335 (2007)
4. Bianconi, F., Fernández, A.: Rotation invariant co-occurrence features based on digital circles and discrete Fourier transform. Pattern Recogn. Lett. **48**, 34–41 (2014)
5. Bianconi, F., Fernández, A.: A unifying framework for LBP and related methods. In: Brahnam, S., Jain, L.C., Nanni, L., Lumini, A. (eds.) Local Binary Patterns: New Variants and Applications. Studies in computational intelligence, vol. 506, pp. 17–46. Springer (2014)
6. Caputo, B., Hayman, E., Mallikarjuna, P.: Class-specific material categorisation. In: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV 2005), vol. II, pp. 1597–1604 (2005)
7. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: Proceedings of the British Machine Vision Conference 2014, Nottingham, United Kingdom, September 2014
8. Cimpoi, M., Maji, S., Vedaldi, A.: Deep filter banks for texture recognition and segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, USA, pp. 3828–3836, June 2015
9. Clarke, A.D.F., Halley, F., Newell, A.J., Griffin, L.D., Chantler, M.J.: Perceptual similarity: a texture challenge. In: Proceedings of the British Machine Vision Conference 2011, Dundee, UK, August–September 2011
10. Cusano, C., Napoletano, P., Schettini, R.: Combining multiple features for color texture classification. J. Electron. Imaging **25**(6) (2016)
11. Cusano, C., Napoletano, P., Schettini, R.: Evaluating color texture descriptors under large variations of controlled lighting conditions. J. Opt. Soc. Am. A **33**(1), 17–30 (2016)
12. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. Syst. Man Cybern. **3**(6), 610–621 (1973)
13. Hayman, E., Caputo, B., Fritz, M., Eklundh, J.-O.: On the significance of real-world conditions for material classification. In: Proceedings of the 8th European Conference on Computer Vision (ECCV 2004), Prague, Czech Republic. LNCS, vol. 3024, pp. 253–266. Springer, May 2004
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, Lake Tahoe, USA, vol. 2, pp. 1097–1105 (2012)

15. The KTH-TIPS and KTH-TIPS2 image databases (2004). http://www.nada.kth.se/cvap/databases/kth-tips/. Last Accessed 17 Oct 2016
16. Kylberg, G.: The Kylberg texture dataset v. 1.0. External report (Blue series) 35, Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, Uppsala, Sweden, September 2011
17. Kylberg, G., Sintorn, I.-M.: On the influence of interpolation method on rotation invariance in texture recognition. EURASIP J. Image Video Process. **2016**(1) (2016)
18. Kylberg Sintorn Rotation dataset (2013). http://www.cb.uu.se/~gustaf/KylbergSintornRotation/. Last Accessed 17 Oct 2016
19. Liu, L., Fieguth, P., Wang, X., Pietikäinen, M., Hu, D.: Evaluation of LBP and Deep Texture Descriptors with a New Robustness Benchmark. In: Proceedings of the 14th European Conference on Computer Vision (ECCV 2016). LNCS, Amsterdam, The Netherlands, vol. 9907, pp. 69–86. Springer (2016)
20. MatConvNet: CNNs for MATLAB (2016). http://www.vlfeat.org/matconvnet/. Last Accessed 25 Oct 2016
21. MondialMarmi: A collection of images of polished natural stones for colour and texture analysis. version 2.0 (2015). http://dismac.dii.unipg.it/mm. Last Accessed 17 Oct 2016
22. Ojala, T., Pietikäinen, M., Mäenpää, T., Viertola, J., Kyllönen, J., Huovinen, S.: Outex - new framework for empirical evaluation of texture analysis algorithms. In: Proceedings of the 16th International Conference on Pattern Recognition (ICPR 2002), Quebec, Canada, vol. 1, pp. 701–706. IEEE Computer Society (2002)
23. Parkhi, O.M., Vedaldi, A., Zissermann, A.: Deep face recognition. In: Proceedings of the British Machine Vision Conference 2015, Swansea, United Kingdom, September 2015
24. Pertex database (2011). http://www.macs.hw.ac.uk/texturelab/resources/databases/pertex/. Last Accessed 17 Oct 2016
25. RawFooT, D.B.: Raw food texture database (2015). http://projects.ivl.disco.unimib.it/rawfoot/. Last Accessed 17 Oct 2016
26. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Columbus, USA, pp. 512–519, June 2014
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556 (2014)
28. Vedaldi, A., Lenc, K.: MatConvNet: convolutional neural networks for MATLAB. In: MM 2015 - Proceedings of the 2015 ACM Multimedia Conference, Brisbane, Australia, pp. 689–692, October 2015
29. Xie, X., Mirmehdi, M.: A galaxy of texture features. In: Mirmehdi, M., Xie, X., Suri, J. (eds.) Handbook of Texture Analysis, pp. 375–406. Imperial College Press (2008)
30. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. Int. J. Comput. Vision **73**(2), 213–238 (2007)
31. Zhong, Y., Sullivan, J., Li, H.: Face attribute prediction using off-the-shelf CNN features. In Proocedings of the 2016 International Conference on Biometrics, ICB 2016, Halmstad, Sweden, 6 2016

Images Selection and Best Descriptor Combination for Multi-shot Person Re-identification

Yousra Hadj Hassen(✉), Kais Loukil, Tarek Ouni, and Mohamed Jallouli

National School of Engineers of Sfax, Computer and Embedded Systems Laboratory, University of Sfax, 4.7 km Street of Soukra, 3038 Sfax, Tunisia
hadjhassen.yousra@gmail.com
http://www.ceslab.org
http://www.enis.rnu.tn

Abstract. To re-identify a person is to check if he/she has been already seen over a cameras network. Recently, re-identifying people over large public cameras networks has become a crucial task of great importance to ensure public security. The vision community has deeply studied this area of research. Most existing researches rely only on the spatial appearance information extracted from either one (single-shot) or multiple images (multi-shot) for each person. Actually, the real person re-identification framework is a multi-shot scenario. However, to efficiently model a person's appearance and to select the most informative samples remain a challenging problem. In this work, an extensive comparison of descriptors of state of art associated to the proposed frame selection method is considered. Specifically, we evaluate the samples selection approach using different known descriptors. For fair comparisons, two standard datasets PRID 2011 and iLIDS-VID are used showing the effectiveness and advantages of the proposed method.

Keywords: Camera network · Descriptor · Model · Multi-shot · Person re-identification · Selection

1 Introduction

Recently, person re-identification (re-id) in non-overlapped cameras network presents a crucial task for many real applications like video surveillance, multimedia applications, behavior recognition,... [1]. To re-identify a person is to match his identity across camera views despite the changes that may occur. Actually, many environmental constraints can alter a persons appearance over different cameras views such as luminance variations, different point of view, scale zooming as shown in Fig. 1.

The proposed approaches can be classified either as single-shot or multiple-shot methods depending on the number of images used to construct the person identity. Contrary to the use of a single image to re-identify a person, the multi-shot based methods have been largely studied and significant results are achieved

G. De Pietro et al. (eds.), *Intelligent Interactive Multimedia Systems and Services 2017*,
Smart Innovation, Systems and Technologies 76, DOI 10.1007/978-3-319-59480-4_2

[2]. They mainly focus on designing discriminative feature descriptors collected over many images of the same person. Whereas, real objectives of person re-id steel far from being reached because both relatively reduced execution time and robust feature descriptor are required.

In this paper, the trade-off between the use of robust descriptor and the reduction of execution time is considered. Multiple-shot proposed methods give potential results for person re-id while ignoring the selection of shots used for the re-id. Actually, most of multiple-shot methods select randomly the images forming the identity of the person but try to generate sufficiently robust descriptors that handle appearance changes caused by scale, lighting variations, view angles conditions and occlusions (Fig. 1). Since, for the multi-shot case, the results of images selection methods have a strong impact not only on the descriptor used for re-id, but also on the overall processing time of the system, the selection of both robust descriptor and discriminative frames to construct representative identity for each person is studied.

Fig. 1. Re-identification system constraints.

A key frame selection method is proposed and a rich comparison of recent robust proposed descriptors is associated to find the most performing combination for a real person re-id system.

The paper is organized in 5 sections. The first section is the introduction, followed by Sect. 2 of the related work. Section 3 describes the overall framework. The evaluation and the results comparison are detailed in Sect. 4. The final section is the conclusion and the future works.

2 Related Work

The contribution of this work may be considered on two fields; person re-identification descriptors as well as samples selection methods.

2.1 Descriptors

In person description, the most commonly used features are color and texture. Symmetry Driven Accumulation of Local Features (SDALF) [3] exploits the symmetric property of a person through obtaining head, torso, and leg positions to handle view variations. Gheissari et al. [4] propose a spatial-temporal segmentation method to detect stable foreground regions. For a local region, an HS histogram and an edgel histogram are computed. The latter encodes the dominant local boundary orientation and the RGB ratios on either sides of the edgel.

Hirzer et al. propose a generic descriptive statistical model in [5] and the appearance is modeled by a set of region covariance descriptors. Gray and Tao [6] use 8 color channels (RGB, HS, and YCbCr) and 21 texture filters on the luminance channel, and the pedestrian is partitioned into horizontal stripes. Similarly, Mignon et al. [7] build the feature vector from RGB, YUV and HSV channels and the LBP texture histograms in horizontal stripes. In [8–12], the 32-dim LAB color histogram and the 128-dim SIFT descriptor are extracted from each 10×10 patch densely sampled with a step size of 5 pixels. Das et al. [13] apply HSV histograms on the head, torso and legs from the silhouette. Pedagadi et al. [14] extract color histograms and moments from HSV and YUV spaces before dimension reduction using PCA. Liu et al. [15] extract the HSV histogram, gradient histogram and the LBP histogram for each local patch. In [16], Liao et al. propose the local maximal occurrence (LOMO) descriptor, which includes the color and SILTP histograms. Bins in the same horizontal stripe undergo max pooling and a three-scale pyramid model is built before a log transformation.

Ayedi et al. propose a multi-scale covariance descriptor in [17] using a quad-tree feature to tackle scale zooming appearance and occlusions in person re-id. In [18], Zheng et al. propose extracting the 11-dim color names descriptor for each local patch, and aggregating them into a global vector through a Bag-of-Words (BoW) model. In [19], a hierarchical Gaussian feature is proposed to describe color and texture cues, which models each region by multiple Gaussian distributions. Each distribution represents a patch inside the region. Liu et al. [20] improve the latent Dirichlet allocation (LDA) model using annotated attributes to filter out noisy LDA topics. In [21], Su et al. embed the binary semantic attributes of the same person but different cameras into a continuous low-rank attribute space, so that the attribute vector is more discriminative for matching. Shi et al. [22] propose learning a number of attributes including color, texture, and category labels from existing fashion photography. In [23], where the image is divided into a number of subblocks, each with its associated color histogram, multi-precision similarity matching is granted despite scale and lighting variations.

2.2 Sample Selection

Video summarization is the most known field in representative selection applications. The problem has been treated using clustering, vector quantization [24–26] or sparse selection [27–29]. In [29], the Sparse Modeling Representative Selection (SMRS), based on the summary of videos by considering the proximity of the selected frames in the timeline, removes redundant frames. Intra and inter iteration redundancy splitting is proposed in [30] and significant re-id results are achieved.

Most of proposed representative samples selection methods rely on appearance variation in time, however, time information is almost unavailable in person re-id datasets. The proposed framework, based on key clusters and key frames selection, takes care of this issue. This enables the selection of as many informative samples as possible to improve the identification performance but at the same time avoids tedious training and useless gallery images. Multi-shot are outperforming single-shot re-id approaches. Incorporating supervision using training data leads to superior performance, which is the goal of metric learning. However, for a multi-shot metric learning person re-id approach with superior real scenario performance, the amount of data presents a hard issue for training time. Thats why, we propose a multi-shot metric learning person re-id framework based on continuous representative samples selection.

3 Proposed Approach

In Fig. 2, the input of the proposed framework is a large gallery of images mostly formed by large sequence of frames of different persons captured in one camera view.

The proposed overall scheme of person re-id is composed of three main parts; first, features are extracted from the images in order to project the visual appearance into concrete parameters (color, texture, position, pose view). After that, a key frame selection algorithm is introduced [31]. The goal is to keep only informative images for each person so that all the informative appearance variations over time and space are summarized and useless noisy and redundant frames are removed. The major contribution of this work is to get the best combination of the descriptor (set of features) and the key frame selection algorithm so that a relatively speed and robust (accurate) re-id is achieved. To that end, the inter and intra descriptors distances are computed. The feed-back enables the evaluation of the used features and the algorithm converges to the best (descriptor, key frame selection) combination. Thus, a new gallery is formed containing key frames descriptors. Finally, the matching block allows the mapping of a newly unknown observed individual in another camera view (a probe) to one of the identity stored in the gallery yet constructed.

3.1 Features Extraction

As detailed in section two, different descriptors are proposed and excellent performances are reached. It is pretty promising to construct robust identities for each person. However, execution time and memory consumption present crucial constraints that must be treated for real video surveillance applications. Therefore,