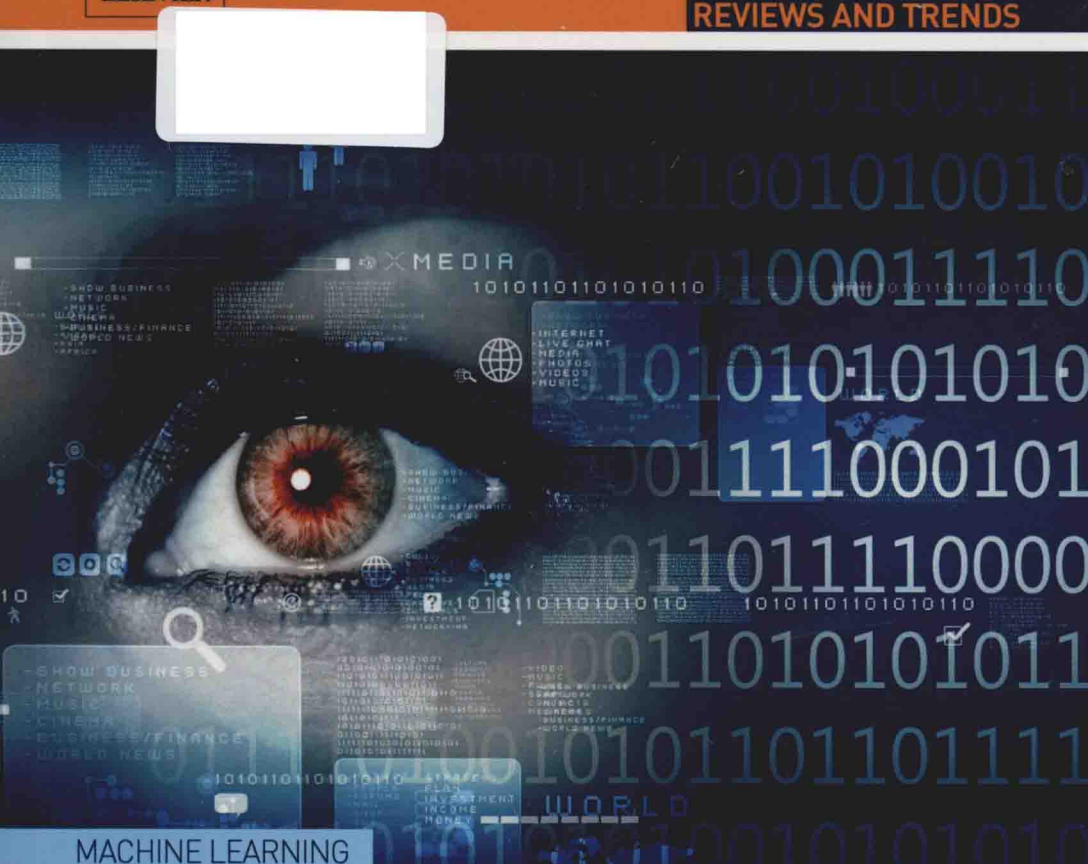




COMPUTER SCIENCE
REVIEWS AND TRENDS



LEARNING-BASED LOCAL VISUAL REPRESENTATION AND INDEXING

RONGRONG JI | YUE GAO | LING-YU DUAN | HONGXUN YAO | QIONGHAI DAI

Learning-Based Local Visual Representation and Indexing

By

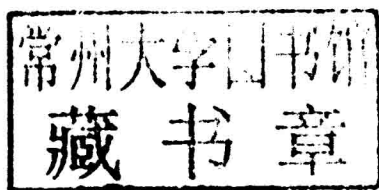
Rongrong Ji

Yue Gao

Ling-Yu Duan

Hongxun Yao

Qionghai Dai



Amsterdam • Boston • Heidelberg • London • New York • Oxford
ELSEVIER Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Executive Editor: Steve Elliot
Editorial Project Manager: Lindsay Lawrence
Project Manager: Anusha Sambamoorthy
Designer: Matthew Limbert

Elsevier

Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
225 Wyman Street, Waltham, MA 02451, USA
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

Copyright © 2015 Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN: 978-0-12-802409-6

For information on all Elsevier publications
visit our website at store.elsevier.com

This book has been manufactured using Print On Demand technology. Each copy is produced to order and is limited to black ink. The online version of this book will show color figures where appropriate.



**Working together
to grow libraries in
developing countries**

www.elsevier.com • www.bookaid.org

Learning-Based Local Visual Representation and Indexing

PREFACE

The visual local representation model based on local features and visual vocabulary serves as a fundamental component in many existing computer vision systems. It has widespread application in the fields of object recognition, scene matching, multimedia content search and analysis, and also is the ad hoc focus of current computer vision and multimedia analysis research. The pipeline of the visual local representation model is to first extract the local interest points from images, then quantize such points into visual vocabulary, which forms a quantization table to obtain the feature-space division into visual words. Subsequently, each image is represented as a bag-of-visual-words descriptor, and is inverted indexed into all its corresponding visual words. Research on current computer vision systems have shown that local visual representation models have sufficient robustness against scale and affine transforms and are good at handling partial object occlusion and matching.

However, recent research has also discovered that there are problems in the state-of-the-art visual local representation models, i.e., insufficient visual content discriminability, extreme dense representation, as well as an inability to reveal higher-level semantics. This book focuses on the study of local feature extraction, quantization errors and semantic discriminability in visual vocabulary, as well as the visual quantization errors, semantic discriminability during the visual vocabulary construction, and the visual phrase based visual vocabulary representation problem.

In the local feature extraction, both spatial and category contexts are exploited, which puts forward the interest-point detection from a local scope toward a global scope. In the unsupervised learning of visual vocabulary and its indexing, the quantization errors in the traditional visual vocabulary are investigated, which further reveals the difference between visual words and textual words, and the influence of narrowing this difference. In the supervised learning of visual vocabulary and its indexing, the image labels are introduced to supervise the visual vocabulary construction, which achieves learning-based quantization in local feature space. Finally, based on the optimized visual vocabulary model, the extension from visual words

to visual phrases is investigated, together with its usage and combination manners with the traditional bag-of-visual-words representation. The main contents of this book are as follows.

In the stage of interest-point detection, a context-aware semi-local interest-point detector is proposed. This detector integrates maximum outputs in image scale space with spatial correlations for interest-point detection. First, the multiple-scale spatial correlations of local features are integrated into a difference of contextual Gaussian (DoCG) field. Experiments have revealed that it can fit the global saliency analysis results to a certain degree. Second, the mean shift algorithm is adopted to locate the detection results within the difference of contextual Gaussian field, in which the training labels are also integrated into the mean shift kernels to enable the finding of “interest” points for subsequent classifier training.

In the stage of unsupervised learning for constructing visual vocabulary and its indexing, a density-based metric learning is proposed for unsupervised quantization optimization. First, using fine quantization in informative feature space and coarse quantization in uninformative feature space, the quantization errors in visual vocabulary construction are minimized, which produces more similar distribution from visual words to textual words. Second, a boosting chain-based hierarchical recognition and voting scheme is proposed, which improves the online recognition efficiency while maintaining its effectiveness and discriminability.

In the state of supervised visual vocabulary learning, a semantic embedding-based supervised quantization approach is proposed. This approach introduces the image labels from the web to build the semantic sensitive visual vocabulary. First, a feature-space density-diversity estimation algorithm is introduced to propagate the image labels from image level into local feature level. Second, the supervised visual vocabulary construction is modeled into a hidden Markov random field, in which the observed field models the local feature set, while the hidden field models the user label supervision. The supervision in the hidden field is achieved via Gibbs distribution over the observed field, and the vocabulary construction is treated as a supervised clustering procedure on the observed field. Meanwhile, we adopt WordNet to model the semantic correlations for user labels in the hidden field, which effectively eliminates the labels synonym.

In the stage of visual vocabulary-based representation, a co-location visual pattern mining algorithm is proposed. This algorithm encodes the spatial co-occurrence and correlative positions of local feature descriptors into co-location transactions and leverages Apriori algorithm to mine the co-location visual patterns. Such a pattern is second order and is sensitive to category information, which serves as more discriminative and lower dimensional local visual descriptions. In addition, such sparse representation, together with the original bag-of-visual-words representation, can further improve the visual search precision in visual search and recognition experiments in benchmark databases, which has been proven in quantitative experimental comparisons.

CONTENTS

Preface	vii
List of Figures	xi
List of Tables.....	xv
List of Algorithms	xvii
Chapter 1 Introduction	1
1.1 Background and Significance	1
1.2 Literature Review of the Visual Dictionary	4
1.3 Contents of This Book	14
Chapter 2 Interest-Point Detection: Beyond Local Scale.....	17
2.1 Introduction.....	17
2.2 Difference of Contextual Gaussians	20
2.3 Mean Shift-Based Localization	22
2.4 Detector Learning	25
2.5 Experiments	28
2.6 Summary	39
Chapter 3 Unsupervised Dictionary Optimization	41
3.1 Introduction	41
3.2 Density-Based Metric Learning.....	44
3.3 Chain-Structure Recognition	48
3.4 Dictionary Transfer Learning.....	51
3.5 Experiments	54
3.6 Summary	65

Chapter 4 Supervised Dictionary Learning via Semantic Embedding 67

4.1 Introduction 67

4.2 Semantic Labeling Propagation 67

4.3 Supervised Dictionary Learning 70

4.4 Experiments 75

4.5 Summary 80

Chapter 5 Visual Pattern Mining 81

5.1 Introduction 81

5.2 Discriminative 3D Pattern Mining 84

5.3 CBoP for Low Bit Rate Mobile Visual Search 92

5.4 Quantitative Results 93

5.5 Conclusion 98

Conclusions 101

References 103

LIST OF FIGURES

1.1	Visualized example of a visual dictionary.	10
2.1	Influences of different contextual and Mean Shift scales.	23
2.2	Proposed descriptor for CASL detectors.....	24
2.3	Results of learning-based CASL detection.....	26
2.4	Examples from the UKBench retrieval benchmark database.	28
2.5	Examples from the Caltech101 object categorization database.....	29
2.6	Repeatability comparison in detector repeatability sequence.....	32
2.7	CASL performance comparison in near-duplicated image retrieval.	34
2.8	Categorization confusion matrix in 10 categories from Caltech101 (I).	36
2.9	Categorization confusion matrix in 10 categories from Caltech101 (II).	36
3.1	Visual word distribution in a 2-layer, 12-level vocabulary tree.	43
3.2	Feature-Frequency statistics (the scale of each axis is given by Log-Log).	46
3.3	Hierarchical TF-IDF term weighting. Each hierarchial level is treated as a higher-level “visual word”.....	49
3.4	Hierarchical recognition chain by a vocabulary tree.....	50
3.5	Exemplar photos in SCity database.	56
3.6	Vocabulary tree-based visual recognition model flowchart.	56
3.7	Performance comparison using original vocabulary tree in UKBench.	57
3.8	Performance comparison using DML-based vocabulary tree in UKBench.	58
3.9	Performance comparison using original vocabulary tree in SCity.	58

3.10	Performance comparison using DML-based vocabulary tree in SCity.	59
3.11	Visual words distribution in 1-way, 12-layer dictionary in SCity.	59
3.12	Visualized results of quantization error reduction.	60
3.13	Precision and time comparison between hierarchical recognition chain (1-way) and GNP. (GNP number: 1-11).	60
3.14	Performance of hierarchical chain at different hierarchy levels.	61
3.15	Performance of dictionary transfer learning from SCity to UKBench.	61
3.16	Dictionary transfer performance from UKBench to SCity.	62
3.17	Recognition model updating.	62
3.18	Sequential indexing without trigger criteria.	63
3.19	Time cost with/without trigger criteria.	64
3.20	Incremental indexing with trigger criterion.	65
4.1	Semantic embedding framework.	68
4.2	Original patch set (partial) and its DDE filtering for “Face”.	70
4.3	Semantic embedding by Markov Random Field.	71
4.4	Ratios between inter-class distance and intra-class distance with and without semantic embedding.	76
4.5	MAP comparisons between GSE and VT, GNP in Flickr.	77
4.6	MAP with different embedding cases.	78
4.7	Comparison with adaptive dictionary in Flickr 60,000.	78
4.8	Confusion tables on PASCAL VOC 05 with comparison to Universal Vocabulary Confusion tables on PASCAL VOC 05 in comparison to universal dictionary.	79
5.1	Exemplar illustrations of incorrect 2D neighborhood configurations of visual words, which are caused by either binding words with diverse depth, or binding words from both foreground and background objects, respectively.	82
5.2	The proposed compact bag of patterns (CBoP) descriptor with application to low bit rate mobile visual search.	83

5.3	Visualized examples about the point clouds for visual pattern candidate construction. Exemplar landmark locations are within Peking University.	86
5.4	Case study of the mined patterns between the gravity-based pattern mining and the Euclidean distance-based pattern mining.	89
5.5	The proposed low bit rate mobile visual search framework using CBoP descriptor. Different from previous works in near-duplicate visual search, we emphasize on extremely compact descriptor extraction directly on the mobile end. To achieve zero-latency query delivery, for each query, our CBoP descriptor is typically hundreds of bits. To the best of our knowledge, it is the most compact descriptor with comparable discriminability to the state-of-the-art visual descriptors [16, 112, 113, 125].	92
5.6	Exemplar local patches in the PhotoTourism dataset. Each patch is sampled as 64×64 gray scale with a canonical scale and orientation. For details of how the scale and orientation is established, please refer to [126]. These ground truth correspondences are collected from the structure-from-motion-based point cloud construction, with the back projection of the matched points.	94
5.7	Exemplar photos collected from Flickr and Panoramio to build our 10M landmarks dataset.	94
5.8	Example comparisons in the extreme mobile query scenarios including Occlusive query set, Background cluttered query set, Night query set, and Blurring and shaking query set in the PKUBench dataset.	97
5.9	Compression rate and ranking distortion analysis with comparison to [112, 113, 125] using the ground truth query set.	98

LIST OF TABLES

2.1	Influence of different contextual scales \mathcal{S}_c and Mean Shift scales \mathcal{S}_m	33
2.2	Quantitative comparisons to contextual global features in Caltech5 subset	37
2.3	Quantitative comparisons to Shape Context in Caltech5 subset (classification phase: SVM).....	38
2.4	Time cost comparisons of different contextual scales (\mathcal{S}_c) and mean shift scales (\mathcal{S}_m).....	38
3.1	Hierarchical quantization errors measured using the overlapping rates (%).....	42
3.2	Performance analysis of VT Shift	63
5.1	Time (Second) requirements for CBoP and other alternatives on the mobile end	97

