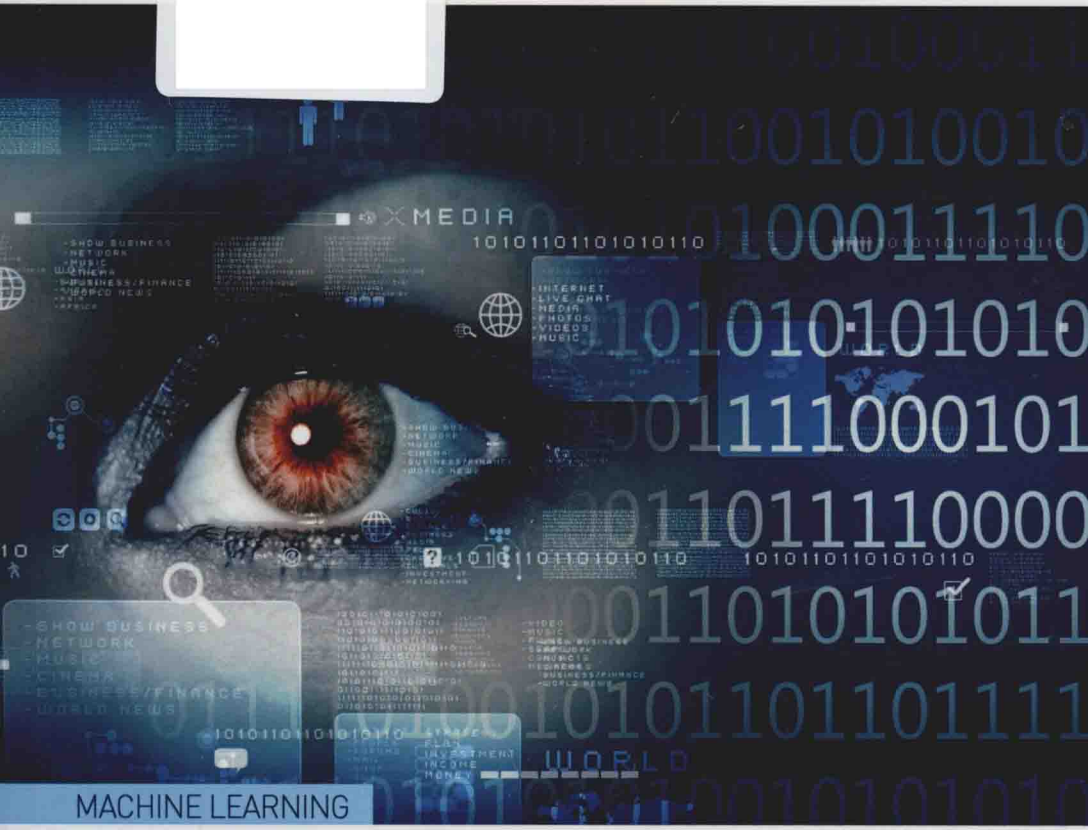
ELSEVIER

COMPUTER SCIENCE
REVIEWS AND TRENDS



MACHINE LEARNING

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

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

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

LIST OF FIGURES

LIST OF TABLES