

The Art of Writing

(technical reports & papers)

Tomas Pajdla

pajdla@cvut.cz

Czech Institute of Informatics, Robotics & Cybernetics
Czech Technical University in Prague





Tomas Pajdla

Associate professor, [Czech Technical University in Prague](#)

Verified email at cvut.cz - [Homepage](#)

[Computer Vision](#) [Robotics](#) [Geometry](#) [Algebraic Geometry](#)

+ FOLLOW

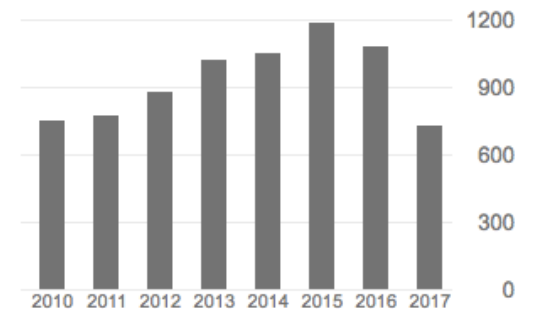
TITLE	CITED BY	YEAR
Robust wide-baseline stereo from maximally stable extremal regions J Matas, O Chum, M Urban, T Pajdla Image and vision computing 22 (10), 761-767	4262	2004
3D with Kinect J Smisek, M Jancosek, T Pajdla Consumer depth cameras for computer vision, 3-25	541	2013
A convenient multicamera self-calibration for virtual environments T Svoboda, D Martinec, T Pajdla PRESENCE: teleoperators and virtual environments 14 (4), 407-422	521	2005
Epipolar geometry for central catadioptric cameras T Svoboda, T Pajdla International Journal of Computer Vision 49 (1), 23-37	239	2002
Robust rotation and translation estimation in multiview reconstruction D Martinec, T Pajdla Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, 1-8	231	2007
Multi-view reconstruction preserving weakly-supported surfaces M Jancosek, T Pajdla Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on ...	214	2011
Epipolar geometry for panoramic cameras T Svoboda, T Pajdla, V Hlaváč Computer Vision—ECCV'98, 218-231	209	1998
Structure from motion with wide circular field of view cameras B Micusik, T Pajdla IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (7), 1135-1149	177	2006

GET MY OWN PROFILE

Cited by

[VIEW ALL](#)

	All	Since 2012
Citations	10895	5980
h-index	43	31
i10-index	103	67



Co-authors

[VIEW ALL](#)

	Zuzana Kukelova Post Doc Researcher, Microsoft ...	>
	Ondrej Chum CMP, CTU in Prague	>
	Jiri Matas Professor, Czech Technical Univ...	>
	Vaclav Hlavac Professor of engineering cybern...	>
	Michal Havlena Vuforia, PTC Vienna, Austria	>

The Art of Writing

(technical reports & papers)

Good scientific writing is not a matter of life and death;
It is much more serious than that.

[R. A. Day]

The Art of Writing

(technical reports & papers)

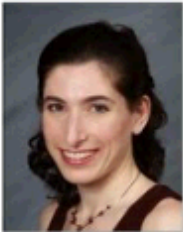
Good scientific writing is a necessary condition to

1. get your work published
2. get your work used
3. be cited
4. receive prizes


It is not enough if the work is not good.


There are (too) many books, lectures, ... about how to write

Watch Kristin!



Write Well and Prosper – Science Writing Tips







If you were writing an essay about lowering the drinking age in the United States, here's some research you might want:


- Facts about drinking age laws in the US (and the history of these laws)
- Statistics about drunk-driving and other accidents caused by drinking
- Quotes from a respected source about drinking

For example...



have built on other scholars' ...
our own conclusions ...

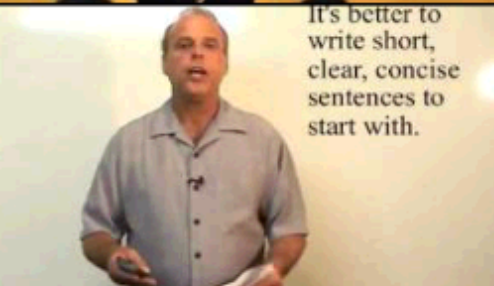




2014/15 Science in Society


Tips for writing a literature review article

It's better to write short, clear, concise sentences to start with.




ACS WEBINARS™
February 3, 2011

Fundamentals of Effective Scientific Writing – Manuscripts and Grants



Thomas Emswiler, Grants Consultancy



Kristin Emswiler, Marquette University


Please submit your questions for the speaker via the Questions Panel in GoToWebinar.

Download slides: <http://acswebinars.org/saman>

Manuscript

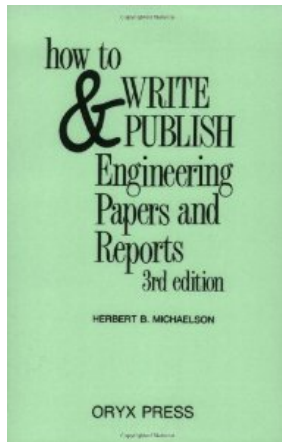
- Read the notes provided for contributors.
- Abstracts should be written in the third person and shouldn't contain references. Abstract writing is a skill, it should NOT be the same as the introduction or the conclusion!
- Ensure references cited in text, appear in bibliography.
- Expand any acronyms, remember it is an international audience.
- Check spelling and grammar carefully.
- ★ Take care when choosing the title, remember academics may find it via a search engine or see it on a content alerting service.
- Always supply keywords if asked.

Continued...



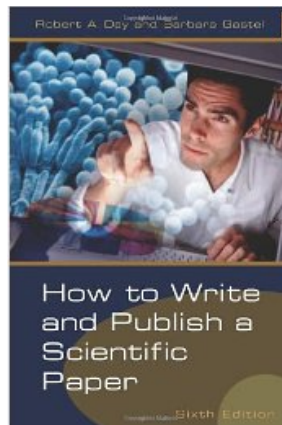


Good “standard” books ...



Herbert B. Michaelson, editor.

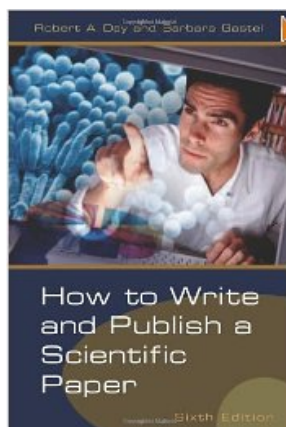
How to Write and Publish Engineering Papers and Reports
Professionals Writing Series. Greenwood , 3rd ed., 1990



Robert A. Day.

How to Write and Publish a Scientific Paper.
Greenwood , 6rd edition, 2006.

... have all the theory about scientific (technical) writing



vi Contents

Chapter 9	
How to Prepare the Abstract	52
Chapter 10	
How to Write the Introduction	56
Chapter 11	
How to Write the Materials and Methods Section	60
Chapter 12	
How to Write the Results	66
Chapter 13	
How to Write the Discussion	69
Chapter 14	
How to State the Acknowledgments	73
Chapter 15	
How to Cite the References	75

Part III: Preparing the Tables and Figures

Chapter 16	
How to Design Effective Tables	85
Chapter 17	
How to Prepare Effective Graphs	92
Chapter 18	
How to Prepare Effective Photographs	99

... but is still may not be enough

since only practice makes perfection.

Anatomy of a technical research paper (an example):

<http://www.ok.ctrl.titech.ac.jp/~torii/project/repttile/download/Torii-CVPR-2013-final.pdf>

Visual Place Recognition with Repetitive Structures

Akihiko Torii
Tokyo Tech*

Josef Sivic
INRIA†

Tomas Pajdla
CTU in Prague‡

Masatoshi Okutomi
Tokyo Tech*

torii@ctrl.titech.ac.jp

Josef.Sivic@inria.fr

pajdla@cmp.felk.cvut.cz

moo@ctrl.titech.ac.jp

Abstract

Repeated structures such as building facades, fences or road markings often represent a significant challenge for place recognition. Repeated structures are notoriously hard for establishing correspondences using multi-view geometry. Even more importantly, they violate the feature independence assumed in the bag-of-visual-words representation which often leads to over-counting evidence and significant degradation of retrieval performance. In this work we show that repeated structures are not a nuisance but, when appropriately represented, they form an important distinguishing feature for many places. We describe a representation of repeated structures suitable for scalable retrieval. It is based on robust detection of repeated image structures and a simple modification of weights in the bag-of-visual-word model. Place recognition results are shown on datasets of street-level imagery from Pittsburgh and San Francisco demonstrating significant gains in recognition performance compared to the standard bag-of-visual-words baseline and more recently proposed burstiness weighting.

1. Introduction

Given a query image of a particular street or a building, we seek to find one or more images in the geotagged database depicting the same place. The ability to visually recognize a place depicted in an image has a range of potential applications including automatic registration of images taken by a mobile phone for augmented reality applications [1] and accurate visual localization for robotics [2]. Scalable place recognition methods [3, 4, 13, 31, 37] often build on the efficient bag-of-visual-words representation developed for object and image retrieval [5, 13, 15, 29, 26, 30]. In an offline pre-processing stage, local invariant descriptors are

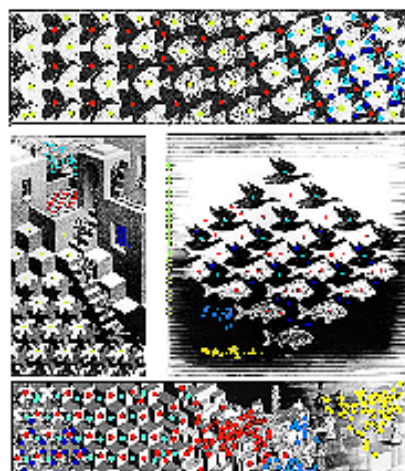


Figure 1. We detect groups of repeated local features (overlaid in colors). The detection is robust against local deformation of the repeated element and makes only weak assumptions on the spatial structure of the repetition. We develop a representation of repeated structures for efficient place recognition based on a simple modification of weights in the bag-of-visual-word model.

extracted from each image in the database and quantized into a pre-computed vocabulary of visual words. Each image is represented by a sparse (weighted) frequency vector of visual words, which can be stored in an efficient inverted file indexing structure. At query time, after the visual words are extracted from the query image, the retrieval proceeds in two steps. First a short-list of ranked candidate images is obtained from the database using the bag-of-visual-words representation. Then, in the second verification stage, candidates are re-ranked based on the spatial layout of visual words.

A number of extensions of this basic architecture have

been proposed. Examples include: (i) learning better visual vocabularies [21, 28]; (ii) developing quantization methods less prone to quantization errors [12, 27, 42]; (iii) combining returns from multiple query images depicting the same scene [3, 4]; (iv) exploiting the 3D or graph structure of the database [11, 20, 29, 43, 43, 47]; or (v) indexing on spatial relations between visual words [39, 47, 48].

In this work we develop a scalable representation for large-scale matching of repeated structures. While repeated structures often occur in man-made environments – examples include building facades, fences, or road markings – they are usually treated as nuisance and downweighted at the indexing stage [13, 18, 36, 39]. In contrast, we develop a simple but efficient representation of repeated structures and demonstrate its benefits for place recognition in urban environments. In detail, we first robustly detect repeated structures in images by finding spatially localized groups of visual words with similar appearance. Next, we modify the weights of the detected repeated visual words in the bag-of-visual-word model, where multiple occurrences of repeated elements in the same image provide a natural soft-assignment of features to visual words. In addition the contribution of repetitive structures is controlled to prevent dominating the matching score.

The rest of the paper is organized as follows. After describing related work on finding and matching repeated structures (Section 1), we review in detail (Section 2) the common tf-idf visual word weighting scheme and its extensions to soft-assignment [27] and repeated structure suppression [13]. In Section 3 we describe our method for detecting repeated visual words in images. In Section 4 we describe the proposed model for scalable matching of repeated structures, and demonstrate its benefits for place recognition in section 5.

Related work. Detecting repeated patterns in images is a well-studied problem. Repetitions are often detected based on an assumption of a single pattern repeated on a 2D (deformed) lattice [10, 19, 23]. Special attention has been paid to detecting planar patterns [33, 38] and in particular building facades [3, 4, 43], for which highly specialized grammar models, learnt from labelled data, were developed [23, 41].

Detecting planar repeated patterns can be useful for single view facade rectification [3] or even single-view 3D reconstruction [46]. However, the local ambiguity of repeated patterns often presents a significant challenge for geometric image matching [33, 38] and image retrieval [13].

Schindler *et al.* [38] detect repeated patterns on building facades and then use the rectified repetition elements together with the spatial layout of the repetition grid to estimate the camera pose of a query image, given a database of building facades. Results are reported on a dataset of 5 query images and 9 building facades. In a similar spirit,

Doubek *et al.* [8] detect the repeated patterns in each image and represent the pattern using a single shift-invariant descriptor of the repeated element together with a simple descriptor of the 2D spatial layout. Their matching method is not scalable as they have to exhaustively compare repeated patterns in all images. In scalable image retrieval, Jegou *et al.* [13] observe that repeated structures violate the feature independence assumption in the bag-of-visual-word model and test several schemes for down-weighting the influence of repeated patterns.

2. Review of visual word weighting strategies

In this section we first review the basic tf-idf weighting scheme proposed in text retrieval [32] and also commonly used for the bag-of-visual-words retrieval and place recognition [3, 8, 12, 13, 18, 23, 26, 40]. Then, we discuss the soft-assignment weighting [27] to reduce quantization errors and the ‘burstiness’ model recently proposed by Jegou *et al.* [13], which explicitly downweights repeated visual words in an image.

Term frequency–inverse document frequency weighting. The standard ‘term frequency–inverse document frequency’ (tf-idf) weighting [32], is computed as follows. Suppose there is a vocabulary of V visual words, then each image is represented by a vector

$$\mathbf{v}_d = (t_1, \dots, t_i, \dots, t_V)^T \quad (1)$$

of weighted visual word frequencies with components

$$t_i = \frac{n_{di}}{n_d} \log \frac{N}{N_i}, \quad (2)$$

where n_{di} is the number of occurrences of visual word i in image d , n_d is the total number of visual words in the image d , N_i is the number of images containing term i , and N is the number of images in the whole database. The weighting is a product of two terms: the *visual word frequency*, n_{di}/n_d , and the *inverse document (image) frequency*, $\log N/N_i$. The word frequency weights words occurring more often in a particular image higher (compared to visual word present/absent), whilst the inverse document frequency downweights visual words that appear often in the database, and therefore do not help to discriminate between different images. At the retrieval stage, images are ranked by the normalized scalar product (cosine of angle)

$$f_d = \frac{\mathbf{v}_q^T \mathbf{v}_d}{\|\mathbf{v}_q\|_2 \|\mathbf{v}_d\|_2} \quad (3)$$

between the query vector \mathbf{v}_q and all image vectors \mathbf{v}_d in the database, where $\|\mathbf{v}\|_2 = \sqrt{\mathbf{v}^T \mathbf{v}}$ is the L_2 norm of \mathbf{v} . When both the query and database vectors are pre-normalized to unit L_2 norm, equation (3) simplifies to the standard scalar product, which can be implemented efficiently using inverted file indexing schemes.

*Department of Mechanical and Control Engineering, Graduate School of Science and Engineering, Tokyo Institute of Technology

†WILLLOW project, Laboratoire d’Informatique de l’École Normale Supérieure, ENS/INRIA/CNRS UMR 8548.

‡Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague

A good title is important

1. to catch attention
2. to get the papers through reviews
3. to have some fun

Video Google: A Text Retrieval Approach to Object Matching in Videos

Josef Sivic and Andrew Zisserman

Robotics Research Group, Department of Engineering Science
University of Oxford, United Kingdom



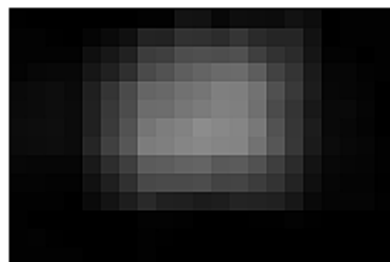
Video Google = Google for videos

All about VLAD

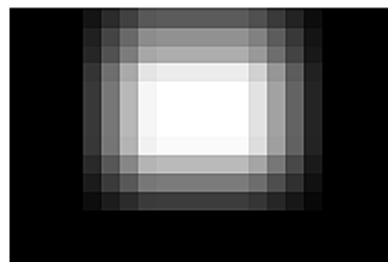
Relja Arandjelović Andrew Zisserman
Department of Engineering Science, University of Oxford
`{relja,az}@robots.ox.ac.uk`



(a) Image with the ROI



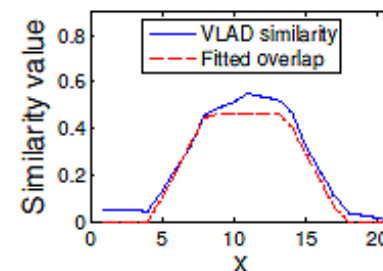
(b) VLAD similarities



(c) Region overlaps



(d) Residuals



(e) 1-D slice

However, one of the most significant contributions in this area has been the introduction of the Vector of Locally Aggregated Descriptors (VLAD) by Jégou *et al.* [8]. This

Tabula Rasa: Model Transfer for Object Category Detection

Yusuf Aytar

Andrew Zisserman

Department of Engineering Science

University of Oxford

{yusuf, az}@robots.ox.ac.uk

Abstract

Our objective is transfer training of a discriminatively trained object category detector, in order to reduce the number of training images required. To this end we propose three transfer learning formulations where a template learnt previously for other categories is used to regularize the training of a new category. All the formulations result in convex optimization problems.

Experiments (on PASCAL VOC) demonstrate significant performance gains by transfer learning from one class to another (e.g. motorbike to bicycle), including one-shot learning, specialization from class to a subordinate class



Figure 1. **The benefit of transfer learning.** The learnt HOG detector template for a motorbike (a) is used as the source for learning a bicycle template together with the samples shown in (b). The resulting learnt bicycle HOG detector template (c) clearly has the shape of a bicycle. Note, here and in the rest of the paper we only visualize the positive components of the HOG vector.

Almost all object category detection methods to date learn the classifier from scratch – tabula rasa. We have proposed a straightforward modification of the learning objective function which retains the benefits of (i) convexity,

Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval

Ondřej Chum¹, James Philbin¹, Josef Sivic¹, Michael Isard² and Andrew Zisserman¹

¹Visual Geometry Group, Department of Engineering Science, University of Oxford

²Microsoft Research, Silicon Valley

{ondra, james, josef, az}@robots.ox.ac.uk

misard@microsoft.com

Abstract

Given a query image of an object, our objective is to retrieve all instances of that object in a large (1M+) image database. We adopt the bag-of-visual-words architecture which has proven successful in achieving high precision at low recall. Unfortunately, feature detection and quantization are noisy processes and this can result in variation in the particular visual words that appear in different images of the same object, leading to missed results.

In the text retrieval literature a standard method for improving performance is query expansion. A number of the highly ranked documents from the original query are reissued as a new query. In this way, additional relevant terms can be added to the query. This is a form of blind relevance feedback and it can fail if ‘outlier’ (false positive) documents are included in the reissued query.

In this paper we bring query expansion into the visual domain via two novel contributions. Firstly, strong spatial constraints between the query image and each result allow us to accurately verify each return, suppressing the false

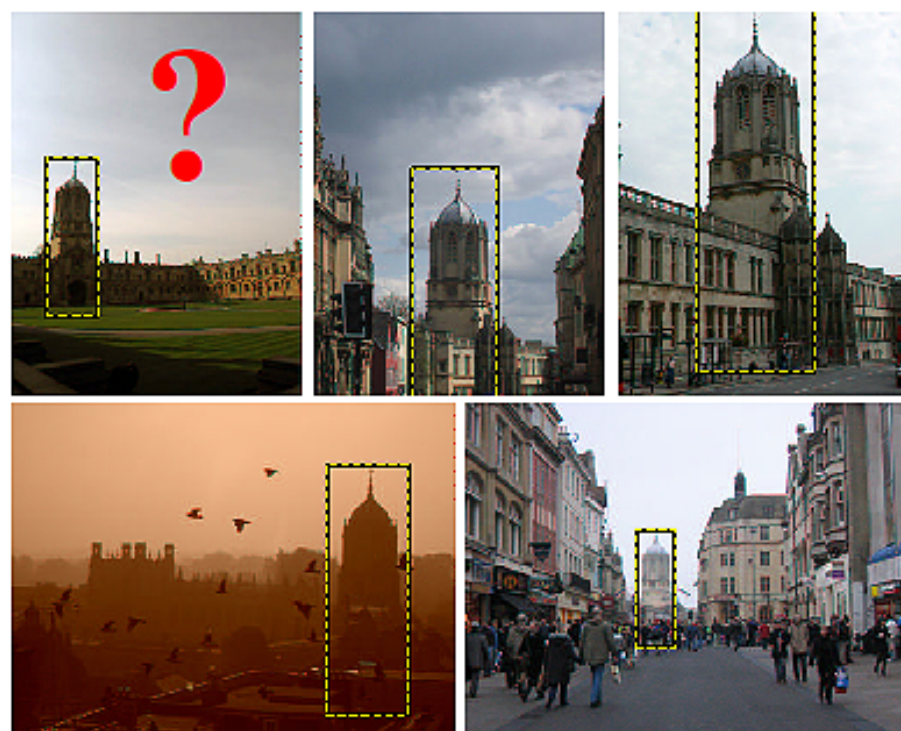


Figure 1. A sample of challenging results returned by our method in answer to a visual query for the *Tom Tower, Christ Church College, Oxford* (top left), which weren't found by a simple bag-of-visual-words method. This query was performed on a large dataset of 1,145,645 images.

A funny title is not everything ...



David Lowe

Professor of Computer Science, University of British Columbia

[Computer Vision - Object Recognition](#)

Verified email at cs.ubc.ca

[Homepage](#)

Citation indices

	All	Since 2008
Citations	42608	31258
h-index	42	35
i10-index	75	54

Citations to my articles



Show: [Next >](#)

[Title / Author](#)

[Cited by](#) [Year](#)

[Distinctive image features from scale-invariant keypoints](#)

DG Lowe

International journal of computer vision 60 (2), 91-110

21295 2004



Paul Viola

Highspot

[Machine Learning](#) - [Computer Vision](#) - [Computational Advertising](#)

Verified email at highspot.com

[Homepage](#)

Citation indices

	All	Since 2008
Citations	29676	19724
h-index	41	32
i10-index	70	58

Citations to my articles



Show: 1-20 [Next >](#)

[Title / Author](#)

[Cited by](#)

[Year](#)

[Rapid object detection using a boosted cascade of simple features](#)

P Viola, M Jones

Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the ...

8348

2001



Josef Sivic

INRIA / Ecole Normale Supérieure

[computer vision](#)

Verified email at ens.fr

[Homepage](#)

Citation indices

	All	Since 2008
Citations	8094	7281
h-index	29	28
i10-index	40	39

Citations to my articles



Show: 1-20 [Next >](#)

[Title / Author](#)

[Cited by](#)

[Year](#)

[Video Google: A text retrieval approach to object matching in videos](#)

J Sivic, A Zisserman

Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on ...

2728

2003

A summary figure

1. tells the story in 1 minute
2. people will remember it
3. ... make it impressive!

Visual Place Recognition with Repetitive Structures

Akihiko Torii
Tokyo Tech*

torii@ctrl.titech.ac.jp

Josef Sivic
INRIA†

Josef.Sivic@ens.fr

Tomas Pajdla
CTU in Prague‡

pajdla@cmp.felk.cvut.cz

Masatoshi Okutomi
Tokyo Tech*

mso@ctrl.titech.ac.jp

Abstract

Repeated structures such as building facades, fences or road markings often represent a significant challenge for place recognition. Repeated structures are notoriously hard for establishing correspondences using multi-view geometry. Even more importantly, they violate the feature independence assumed in the bag-of-visual-words representation which often leads to over-counting evidence and significant degradation of retrieval performance. In this work we show that repeated structures are not a nuisance but, when appropriately represented, they form an important distinguishing feature for many places. We describe a representation of repeated structures suitable for scalable retrieval. It is based on robust detection of repeated image structures and a simple modification of weights in the bag-of-visual-word model. Place recognition results are shown on datasets of street-level imagery from Pittsburgh and San Francisco demonstrating significant gains in recognition performance compared to the standard bag-of-visual-words baseline and more recently proposed business weighting.

1. Introduction

Given a query image of a particular street or a building, we seek to find one or more images in the geotagged database depicting the same place. The ability to visually recognize a place depicted in an image has a range of potential applications including automatic registration of images taken by a mobile phone for augmented reality applications [1] and accurate visual localization for robotics [7]. Scalable place recognition methods [3, 7, 18, 31, 37] often build on the efficient bag-of-visual-words representation developed for object and image retrieval [6, 13, 15, 24, 26, 40]. In an offline pre-processing stage, local invariant descriptors are

*Department of Mechanical and Control Engineering, Graduate School of Science and Engineering, Tokyo Institute of Technology

†WILLow project, Laboratoire d'Informatique de l'École Normale Supérieure, ENSA/NRIA/CNRS UMR 8548.

‡Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague

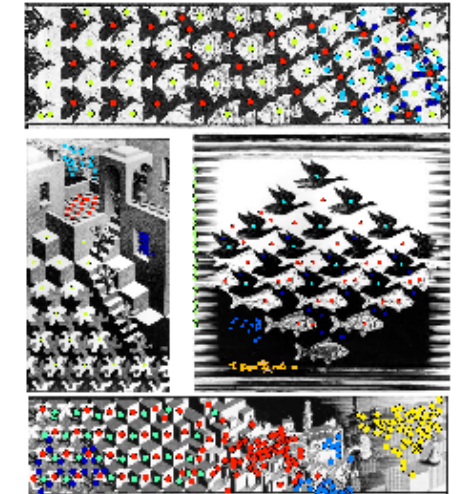


Figure 1. We detect groups of repeated local features (overlaid in colors). The detection is robust against local deformation of the repeated element and makes only weak assumptions on the spatial structure of the repetition. We develop a representation of repeated structures for efficient place recognition based on a simple modification of weights in the bag-of-visual-word model.

extracted from each image in the database and quantized into a pre-computed vocabulary of visual words. Each image is represented by a sparse (weighted) frequency vector of visual words, which can be stored in an efficient inverted file indexing structure. At query time, after the visual words are extracted from the query image, the retrieval proceeds in two steps. First a short-list of ranked candidate images is obtained from the database using the bag-of-visual-words representation. Then, in the second verification stage, candidates are re-ranked based on the spatial layout of visual words.

A number of extensions of this basic architecture have

Abstract

Repeated structures such as building facades, fences or road markings often represent a significant challenge for place recognition. Repeated structures are notoriously hard for establishing correspondences using multi-view geometry. Even more importantly, they violate the feature independence assumed in the bag-of-visual-words representation which often leads to over-counting evidence and significant degradation of retrieval performance. In this work we show that repeated structures are not a nuisance but, when appropriately represented, they form an important distinguishing feature for many places. We describe a representation of repeated structures suitable for scalable retrieval. It is based on robust detection of repeated image structures and a simple modification of weights in the bag-of-visual-word model. Place recognition results are shown on datasets of street-level imagery from Pittsburgh and San Francisco demonstrating significant gains in recognition performance compared to the standard bag-of-visual-words baseline and more recently proposed burstiness weighting.

1. Introduction

Given a query image of a particular street or a building, we seek to find one or more images in the geotagged database depicting the same place. The ability to visually recognize

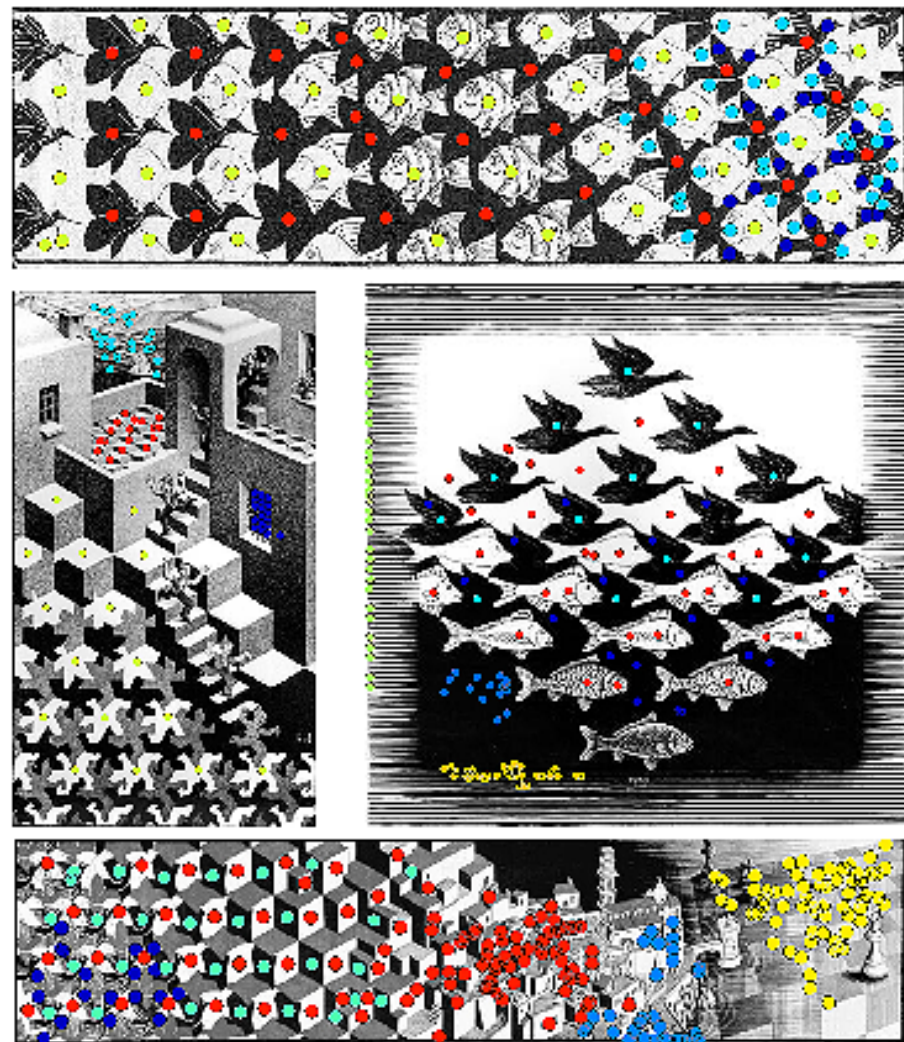


Figure 1. We detect groups of repeated local features (overlaid in colors). The detection is robust against local deformation of the repeated element and makes only weak assumptions on the spatial structure of the repetition. We develop a representation of repeated structures for efficient place recognition based on a simple modification of weights in the bag-of-visual-word model.

Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval

Ondřej Chum¹, James Philbin¹, Josef Sivic¹, Michael Isard² and Andrew Zisserman¹

¹Visual Geometry Group, Department of Engineering Science, University of Oxford

²Microsoft Research, Silicon Valley

{ondra, james, josef, az}@robots.ox.ac.uk

misard@microsoft.com

Abstract

Given a query image of an object, our objective is to retrieve all instances of that object in a large (1M+) image database. We adopt the bag-of-visual-words architecture which has proven successful in achieving high precision at low recall. Unfortunately, feature detection and quantization are noisy processes and this can result in variation in the particular visual words that appear in different images of the same object, leading to missed results.

In the text retrieval literature a standard method for improving performance is query expansion. A number of the highly ranked documents from the original query are reissued as a new query. In this way, additional relevant terms can be added to the query. This is a form of blind relevance feedback and it can fail if ‘outlier’ (false positive) documents are included in the reissued query.

In this paper we bring query expansion into the visual domain via two novel contributions. Firstly, strong spatial constraints between the query image and each result allow us to accurately verify each return, suppressing the false

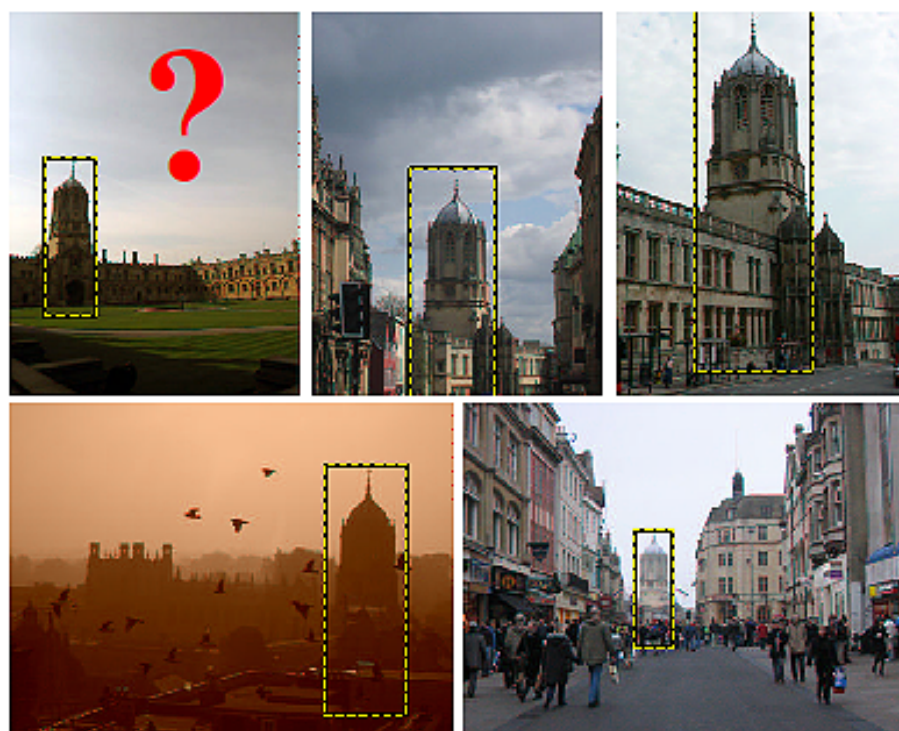


Figure 1. A sample of challenging results returned by our method in answer to a visual query for the *Tom Tower, Christ Church College, Oxford* (top left), which weren't found by a simple bag-of-visual-words method. This query was performed on a large dataset of 1,145,645 images.

Kinectrack: Agile 6-DoF Tracking Using a Projected Dot Pattern

Paul Mollroy*
University of Cambridge
Microsoft Research

Shahram Izadi†
Microsoft Research

Andrew Fitzgibbon‡
Microsoft Research

ABSTRACT

We present Kinectrack, a new six degree-of-freedom (6-DoF) tracker which allows real-time and low-cost pose estimation using only commodity hardware. We decouple the dot pattern emitter and IR camera of the Kinect. Keeping the camera fixed and moving the IR emitter in the environment, we recover the 6-DoF pose of the emitter by matching the observed dot pattern in the field-of-view of the camera to a pre-captured reference image. We propose a novel matching technique to obtain dot pattern correspondences efficiently in wide- and adaptive-baseline scenarios. We also propose an auto-calibration method to obtain the camera intrinsics and dot pattern reference image. The performance of Kinectrack is evaluated and the rotational and translational accuracy of the system is measured relative to ground truth for both planar and multi-planar scene geometry. Our system can simultaneously recover the 6-DoF pose of the device and also recover piecewise planar 3D scene structure, and can be used as a low-cost method for tracking a device without any on-board computation, with small size and only simple electronics.

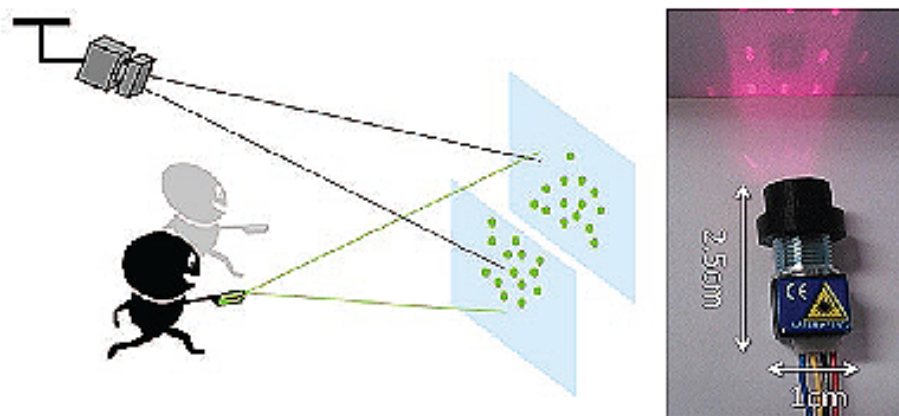


Figure 1: Tracking setup. (Left) A fixed camera, and free moving laser-pattern emitter provides the benefits of “look out” tracking, but the moving device is very simple—a laser, optics, and battery—and can be considerably smaller than “look in” remotes. (Right) The laser and optics of the remote unit. In practice, an infrared laser would be used; a visible-light unit is used here purely for illustration.

KinectFusion: Real-Time Dense Surface Mapping and Tracking*

Richard A. Newcombe
Imperial College London

Shahram Izadi
Microsoft Research

Otmar Hilliges
Microsoft Research

David Molyneaux
Microsoft Research
Lancaster University

David Kim
Microsoft Research
Newcastle University

Andrew J. Davison
Imperial College London

Pushmeet Kohli
Microsoft Research

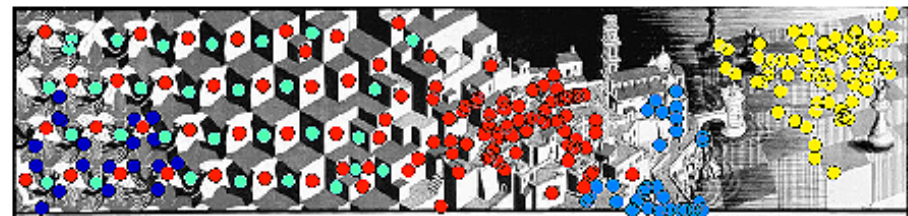
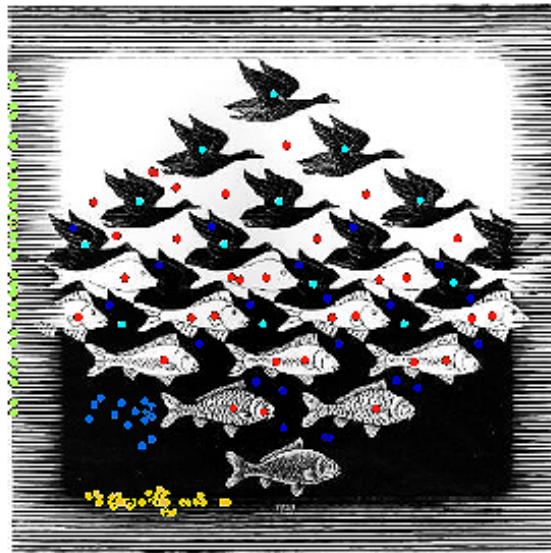
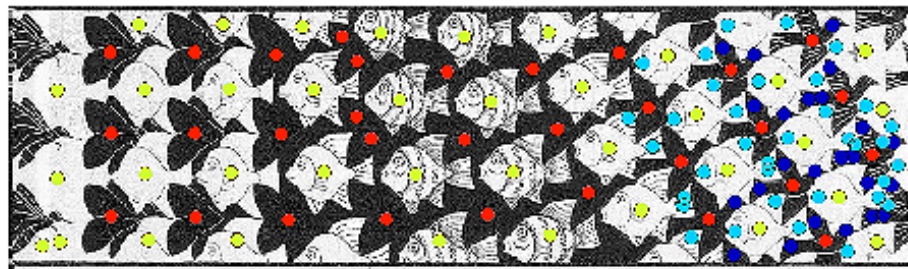
Jamie Shotton
Microsoft Research

Steve Hodges
Microsoft Research

Andrew Fitzgibbon
Microsoft Research



Figure 1: Example output from our system, generated in real-time with a handheld Kinect depth camera and no other sensing infrastructure. Normal maps (colour) and Phong-shaded renderings (greyscale) from our dense reconstruction system are shown. On the left for comparison is an example of the live, incomplete, and noisy data from the Kinect sensor (used as input to our system).



What we do

Figure 1. We detect groups of repeated local features (overlaid in

why it is good

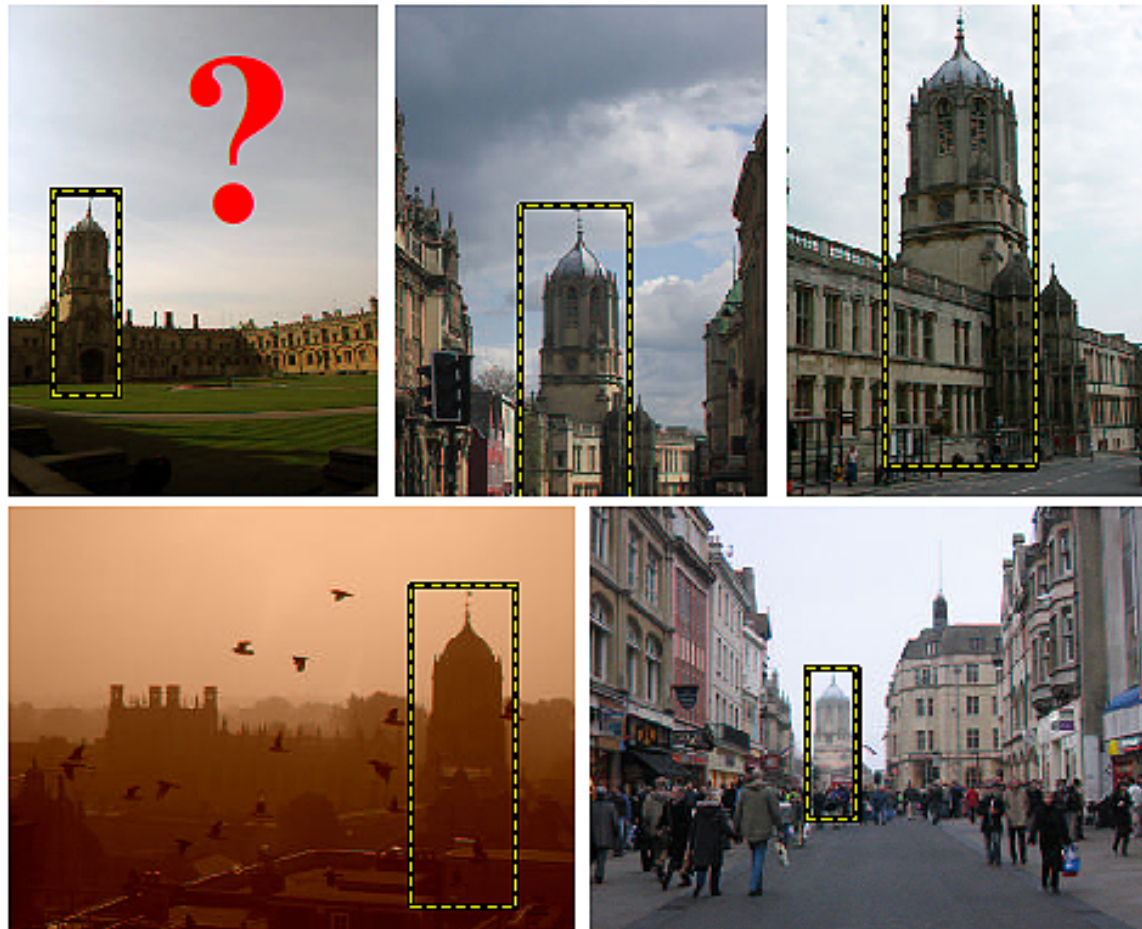
colors). The detection is robust against local deformation of the repeated element and makes only weak assumptions on the spatial structure of the repetition.

what is the interesting detail

We develop a representation of repeated structures for efficient place recognition based on a simple modification of weights in the bag-of-visual-word model.

Figure 1. We detect groups of repeated local features (overlaid in colors). The detection is robust against local deformation of the repeated element and makes only weak assumptions on the spatial structure of the repetition. We develop a representation of repeated structures for efficient place recognition based on a simple modification of weights in the bag-of-visual-word model.

What is the problem?



Great results!

Results

Figure 1. A sample of challenging results returned by our method in answer to a visual query for the *Tom Tower, Christ Church College, Oxford* (top left), which weren't found by a simple bag-of-visual-words method. This query was performed on a large dataset of 1,145,645 images.

Are better than the SoA

And it works on really large data ... it is practical

Abstract

1. Motivation & impact
2. The main contribution (idea, methods, ...)
3. Results/Experiments
4. Single paragraph, no references,

Abstract

Motivation

Repeated structures such as building facades, fences or road markings often represent a significant challenge for place recognition. Repeated structures are notoriously hard for establishing correspondences using multi-view geometry. Even more importantly, they violate the feature independence assumed in the bag-of-visual-words representation which often leads to over-counting evidence and significant degradation of retrieval performance. In this work we show that repeated structures are not a nuisance but, when appropriately represented, they form an important distinguishing feature for many places. We describe a representation of repeated structures suitable for scalable retrieval. It is based on robust detection of repeated image structures and a simple modification of weights in the bag-of-visual-word model. Place recognition results are shown on datasets of street-level imagery from Pittsburgh and San Francisco demonstrating significant gains in recognition performance compared to the standard bag-of-visual-words baseline and more recently proposed burstiness weighting.

Why it is important

Contribution
what is done
&
some details

Results
Experiments
Comparison

Introduction

3 paragraphs

- Motivation & why it is important
- The main contribution compared to the State of the Art
- Structure of the paper

1. Motivation & why it is important

1. Introduction

Given a query image of a particular street or a building, we seek to find one or more images in the geotagged database *depicting the same place*. The ability to visually recognize a place depicted in an image has a range of potential applications including automatic registration of images taken by a mobile phone for augmented reality applications [1] and accurate visual localization for robotics [7]. Scalable place recognition methods [3, 7, 18, 31, 37] often build on the efficient bag-of-visual-words representation developed for object and image retrieval [6, 13, 15, 24, 26, 40]. In an offline pre-processing stage, local invariant descriptors are

What we do

Why it is important

extracted from each image in the database and quantized into a pre-computed vocabulary of visual words. Each image is represented by a sparse (weighted) frequency vector of visual words, which can be stored in an efficient inverted file indexing structure. At query time, after the visual words are extracted from the query image, the retrieval proceeds in two steps. First a short-list of ranked candidate images is obtained from the database using the bag-of-visual-words representation. Then, in the second verification stage, candidates are re-ranked based on the spatial layout of visual words.

been proposed. Examples include: (i) learning better visual vocabularies [21, 28]; (ii) developing quantization methods less prone to quantization errors [14, 27, 44]; (iii) combining returns from multiple query images depicting the same scene [4, 6]; (iv) exploiting the 3D or graph structure of the database [11, 20, 29, 42, 43, 47]; or (v) indexing on spatial relations between visual words [5, 12, 48].

The State of The Art

There must be references:
the more relevant references, the better

2. The main contribution compared to the State of the Art

In this work we develop a scalable representation for large-scale matching of repeated structures. While repeated structures often occur in man-made environments – examples include building facades, fences, or road markings – they are usually treated as nuisance and downweighted at the indexing stage [13, 18, 36, 39]. In contrast, we develop a simple but efficient representation of repeated structures and demonstrate its benefits for place recognition in urban environments. In detail, we first robustly detect repeated structures in images by finding spatially localized groups of visual words with similar appearance. Next, we modify the weights of the detected repeated visual words in the bag-of-visual-word model, where multiple occurrences of repeated elements in the same image provide a *natural soft-assignment* of features to visual words. In addition the contribution of repetitive structures is controlled to prevent dominating the matching score.

The main contribution

Why and how it is new

Technical details

There must be references:
the more relevant references, the better

3. Structure of the paper

The rest of the paper is organized as follows. After describing related work on finding and matching repeated structures (Section 1), we review in detail (Section 2) the common tf-idf visual word weighting scheme and its extensions to soft-assignment [27] and repeated structure suppression [13]. In Section 3 we describe our method for detecting repeated visual words in images. In Section 4, we describe the proposed model for scalable matching of repeated structures, and demonstrate its benefits for place recognition in section 5.

Section 2. Related work in technical detail

Section 3. The new methods

Section 4. The use of the new method

Section 5. Experiments

Section 6. Conclusions

Start every section with a short overview

2. Review of visual word weighting strategies

In this section we first review the basic tf-idf weighting scheme proposed in text retrieval [32] and also commonly used for the bag-of-visual-words retrieval and place recognition [3, 6, 12, 13, 18, 24, 26, 40]. Then, we discuss the soft-assignment weighting [27] to reduce quantization errors and the ‘burstiness’ model recently proposed by Jegou *et al.* [13], which explicitly downweights repeated visual words in an image.

Start every section with a short overview

To give an overview

4. Representing repetitive structures for scalable retrieval

In this section we describe our image representation for efficient indexing taking into account the repetitive patterns. The proposed representation is built on two ideas. First, we aim at representing *the presence* of a repetition, rather than measuring the actual number of matching repeated elements. Second, we note that different occurrences of the same visual element (such as a facade window) are often quantized to different visual words naturally representing

and to provide additional insight

Start every section with a short overview

To state goals

3. Detection of repetitive structures

The goal is to segment local invariant features detected in an image into localized groups of repetitive patterns and a layer of non-repeated features. Examples include detecting repeated patterns of windows on different building facades, as well as fences, road markings or trees in an image (see figure 2). We will operate directly on the extracted local features (rather than using specially designed features [9]) as the detected groups will be used to adjust feature weights in the bag-of-visual-words model for efficient indexing. The feature segmentation problem is posed as finding connected components in a graph.

and to underline the novelty

Equations

The standard ‘term frequency–inverse document frequency’ (*tf-idf*) weighting [32], is computed as follows. Suppose there is a vocabulary of V visual words, then each image is represented by a vector

$$\mathbf{v}_d = (t_1, \dots, t_i, \dots, t_V)^T \quad (1)$$

Number equations

of weighted visual word frequencies with components

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{N_i}, \quad (2)$$

Explain all symbols

where n_{id} is the number of occurrences of visual word i in image d , n_d is the total number of visual words in the image d , N_i is the number of images containing term i , and N is the number of images in the whole database.

Explain what it means

The weighting is a product of two terms: the *visual word frequency*, n_{id}/n_d , and the *inverse document (image) frequency*, $\log N/N_i$. The word frequency weights words occurring more often in a particular image higher (compared to visual word present/absent), whilst the inverse document frequency downweights visual words that appear often in the database, and therefore do not help to discriminate between different images. At the retrieval stage, images are

Figures

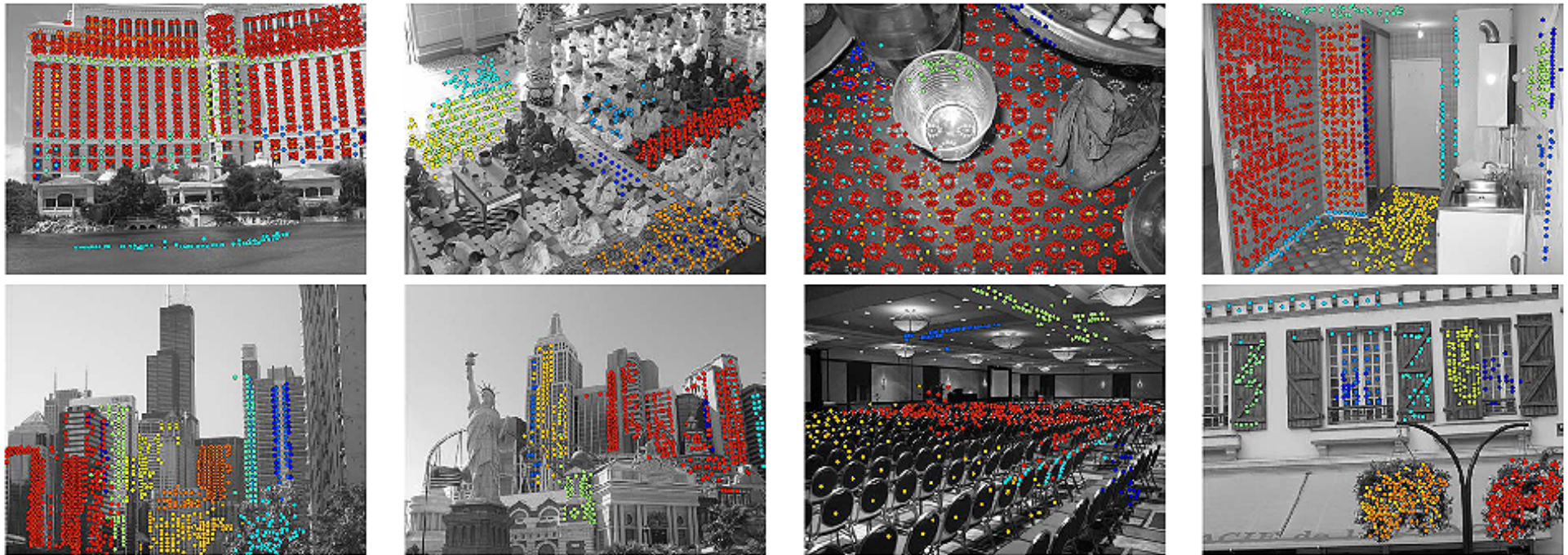
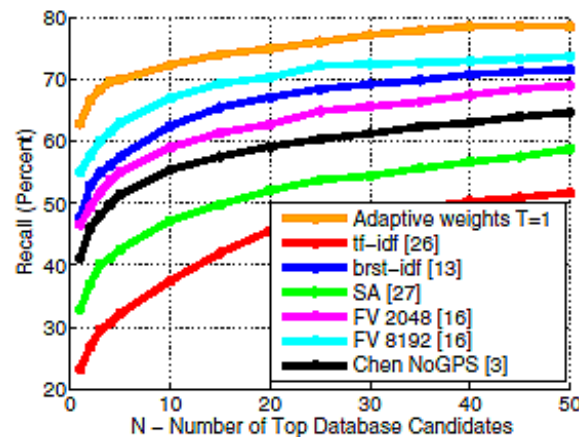


Figure 2. Examples of detected repetitive patterns of local invariant features (“reptiles”) in images from the INRIA Holidays dataset [13]. The different repetitive patterns detected in each image are shown in different colors. The color indicates the number of features in each group (red indicates large and blue indicates small groups). Note the variety of detected repetitive structures such as different building facades, trees, indoor objects, window tiles or floor patterns.

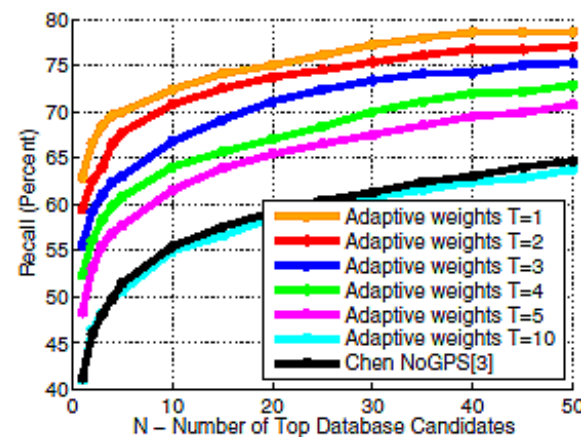
Pay attention to captions:

Title, abstract, figures+captions, equations, references
must give a good overview of the paper in 5-10 mins
(some reviewers read only this)

Figures



(a)



(b)

Figure 6. **Evaluation on the San Francisco [3] dataset.** The fraction of correctly recognized queries (Recall, y-axis) vs. the number of top N retrieved database images (x-axis) for the proposed method (Adaptive weights) compared to several baselines.

Graphs:

1. Describe axes!!!
2. Make it clear (colors, legends, ...)

Experiments

1. Describe the setup, data, ...
2. If necessary, include additional implementation details
3. Present results (graphs, tables, ...)
4. Interpret and compare results. Explain why is your result better (same, worse)

Interpret and compare results

Use graphs to interpret the results
Readers may not see what you want
to demonstrate in the graphs

Next, we evaluate separately the benefits of the two components of the proposed method with respect to the baseline burstiness weights: (i) thresholding using eq. (5) results in +8.92% and (ii) adaptive soft-assignment using eq. (6) and (7) results in +10.30%. When the two are combined the improvement is +11.97%. This is measured for the distance threshold $m = 25$ meters and for the top $N = 10$ but we have observed that the improvements are consistent over a range of N (not shown).

Make clear when you refer to
something that can't be seen in the results

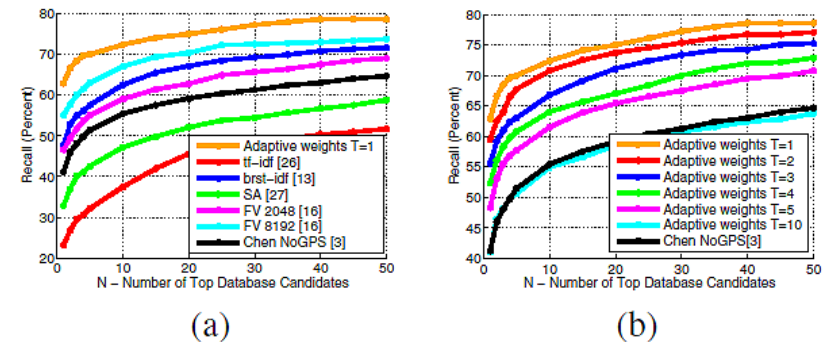


Figure 6. **Evaluation on the San Francisco [3] dataset.** The fraction of correctly recognized queries (Recall, y-axis) vs. the number of top N retrieved database images (x-axis) for the proposed method (Adaptive weights) compared to several baselines.

Conclusion

1 paragraph

- Main contribution (3 times: introduction, main, conclusion)
- No future work (do it but do not write about it)

6. Conclusion

What

In this work we have demonstrated that repeated structures in images are not a nuisance but can form a distinguishing feature for many places. We treat repeated visual words as significant visual events, which can be detected and matched.

How

This is achieved by robustly detecting repeated patterns of visual words in images, and adjusting their weights in the bag-of-visual-word representation. Multiple occurrences of repeated elements are used to provide a natural soft-assignment of features to visual words.

It works
and
has impact

The contribution of repetitive structures is controlled to prevent dominating the matching score. We have shown that the proposed representation achieves consistent improvements in place recognition performance in an urban environment. In addition, the proposed method is simple and can be easily incorporated into existing large scale place recognition architectures.

Acknowledgements

Never forget to acknowledge your sponsors! (and helpers)

Acknowledgements. Supported by JSPS KAKENHI Grant Number 24700161, De-Montes FP7-SME-2011-285839 project, MSR-INRIA laboratory and EIT-ICT labs.

References

1. Cite all technically relevant work but not more
2. Do not use unnecessary details
(try Internet search to make sure it can be found).

References

- [1] B. Aguera y Arcas. Augmented reality using Bing maps., 2010. Talk at TED 2010.
- [2] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.
- [3] D. Chen, G. Baatz, et al. City-scale landmark identification on mobile devices. In *CVPR*, 2011.

Sequence of writing

Paper

1. Title
2. Introduction: the motivation and contribution
3. Figures & captions
4. Equations
5. Previous work
6. The text
7. Abstract
8. Conclusion
9. A better title
10. A better introduction
11. Polish, submit, polish, submit, ...

Sequence of writing

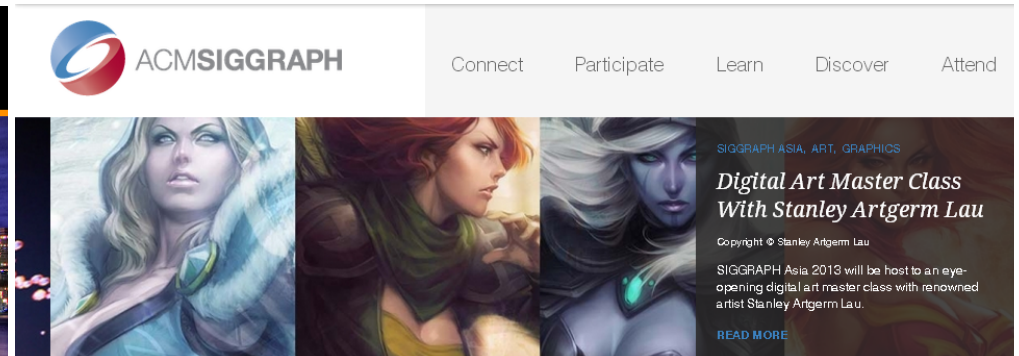
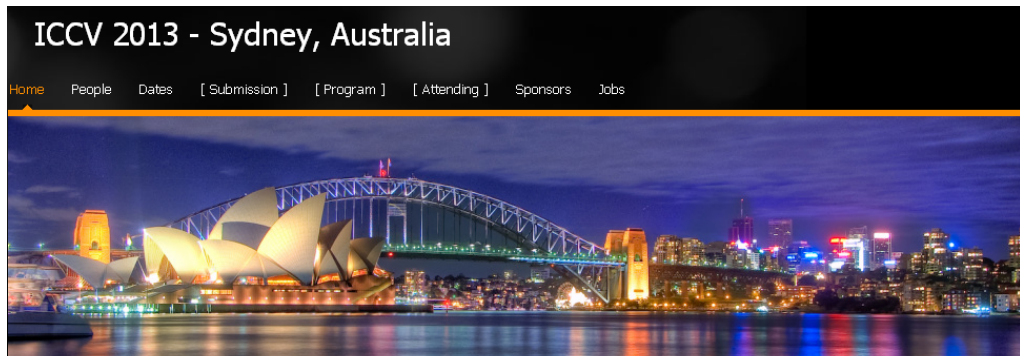
Technical report

1. Main contribution linearly as you work
2. Previous work as is used/compared in the main contribution
3. Introduction
4. Conclusion
5. Abstract

Rebuttal

Paper → reviews → rebuttal → accept/reject decision

Computer vision, machine learning, computer graphics, robotics, ...



Rebuttal

Be very positive

1. We thank the reviewers for their comments. We address the most salient issues below.

Only the most important

2. R1 – “my main concern is that the lack of discussion on why the solver is valid for both planar and non-planar scenes.”

Rephrase the question

It is true that our parameterization is similar to [5] where the non-planar solver degenerates for planar scenes. However, the difference is in the solving method. While the non-planar P4Pfr solver [5] assumes regularity of some matrices, in our new P5Pfr solver we do not have such assumptions and therefore the new solver works for both planar and non-planar scenes. We will add a detailed discussion on this in the paper.

Admit what is correct

Strike back!

Explain how you will improve it

1. R2 – „My concern with this paper is the over-parameterization of the solution when only one radial distortion parameter is used usually allows the noise to affect the solution in a non-wanted way resulting in solutions which are of lower quality than the ones from a closed form solver...”

Shorten where necessary

This is not typically true and we focused on this in the noise experiment in figure 3. The new non-minimal one parameter P5Pfr solver outperforms the minimal P4Pfr solver [5] for all noise levels. In fact using one more point in P5Pfr solver helps better fitting to the noisy data.

Reject criticism politely

Explain the misunderstanding

And add more ...

Practice makes perfection

write, write, write, write, ...

Tomas Pajdla
(pajdla@cmp.felk.cvut.cz)