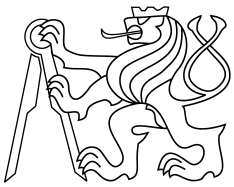
CENTER FOR
MACHINE PERCEPTION



CZECH TECHNICAL
UNIVERSITY IN PRAGUE

MASTER'S THESIS

ISSN 1213-2365

Automated Planar Rectification from Repeated Patterns

James Pritts

prittjam@cmp.felk.cvut.cz

CTU–CMP–2013–04

January 10, 2013

Available at
http://cmp.felk.cvut.cz/~prittjam/pub/mscthesis.pdf

Thesis Advisor: Ondřej Chum

The support of grant GACR P103/12/2310 from The Czech Science Foundation and grant SGS11/125/OHK3/2T/13 from The Grant Agency of The Czech Techinical Univeristy in Prague is gratefully acknowledged.

Research Reports of CMP, Czech Technical University in Prague, No. 4, 2013

Published by

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: http://cmp.felk.cvut.cz

Automated Planar Rectification from Repeated Patterns

James Pritts

January 10, 2013

Czech Technical University in Prague
Faculty of Electrical Engineering

Department of Cybernetics

DIPLOMA THESIS ASSIGNMENT

Student: James Brandon P r i t t s

Study programme: Open Informatics

Specialisation: Computer Vision and Image Processing

Title of Diploma Thesis: Automatic Planar Rectification from Repeated Patterns

Guidelines:

1. Study literature on repeated structure detection, its analysis, and on methods of detecting lines at infinity.
2. Design methods for line at infinity detection from repeated planar structures. Use the geometric information from repeated structures in a single image to rectify the plane to a fronto-parallel view. The methods should not require known camera calibration not known radial distoriton of the camera lens.
3. Implemet the designed algorithm, maximize the precision (use non-linear optimization).
4. Statistically analyse the results.
5. Discuss the failure cases and possible steps to avoid those.

Bibliography/Sources: Will be provided by the supervisor.

Diploma Thesis Supervisor: Mgr. Ondřej Chum, Ph.D.

Valid until: the end of the winter semester of academic year 2013/2014

L.S.

prof. Ing. Vladimír Mařík, DrSc.
Head of Department

prof. Ing. Pavel Ripka, CSc.
Dean

Prague, September 14, 2012

České vysoké učení technické v Praze
Fakulta elektrotechnická

Katedra kybernetiky

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student: James Brandon P r i t t s

Studijní program: Otevřená informatika (magisterský)

Obor: Počítačové vidění a digitální obraz

Název tématu: Automatická rovinná rektifikace z opakujících se vzorů

Pokyny pro vypracování:

1. Seznamte se s literaturou zabývající se detekcí opakujících se struktur, jejich analýzou a s metodami hledání přímek v nekonečnu.
2. Navrhněte metody umožnující nalezení přímek v nekonečnu z rovinného opakujícího se vzoru. Využijte geometrickou informaci v jednom obraze k rektifikaci roviny do fronto-paralelního pohledu pro různé druhy kamer (neznámá kalibrace, radiální zkreslení,...).
3. Implementujte algoritmy, maximalizujte přesnost (použijte nelineární optimalizaci).
4. Výsledky statisticky zhodnoťte.
5. Diskutujte selhání algoritmů a kroky k jejich odstranění.

Seznam odborné literatury: Dodá vedoucí práce.

Vedoucí diplomové práce: Mgr. Ondřej Chum, Ph.D.

Platnost zadání: do konce zimního semestru 2013/2014

prof. Ing. Vladimír Mařík, DrSc.
vedoucí katedry

L.S.

prof. Ing. Pavel Ripka, CSc.
děkan

V Praze dne 14. 9. 2012

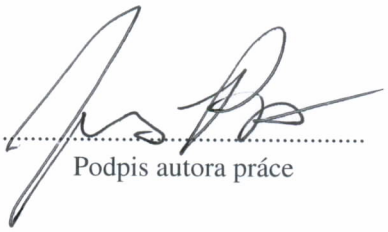
I would like to thank Professor Jiří Matas for the opportunity to be a part of his research group at “G-2”, and for his continued encouragement and support. I’d like to thank Associate Professor Ondřej Chum, who patiently mentored me; provided guidance for my research; and, in particular, supervised this thesis. Also, I am indebted to Dr. Michal Perdoch, who tirelessly assisted me with uncountably many technical issues and sorted many bureaucratic entanglements. Thanks to Karel Lebeda and Karel Lenc for assisting with the Czech translation of the abstract and with typesetting snafus. I would like to express my gratitude to the excellent faculty and staff at Center for Machine Perception for providing an environment that has enabled me to improve as a computer-vision practitioner.

Prohlášení autora práce

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

I hereby declare that this thesis is entirely the result of my own work and all the sources I used are in the list of references in accordance with the Methodological Instructions on Ethical Principles in the Preparation of University Theses.

V Praze dne 3.1.2013



.....

Podpis autora práce

Abstract

Repetitive patterns typically arise from man-made objects and are ubiquitous in image collections. Because of their highly repetitive structure, imaged patterns violate statistical assumptions typically made in scene understanding algorithms, often negatively impacting algorithm efficacy. Conversely, repetitive patterns could be detected and modeled so that their distinctive structures are leveraged to uniquely characterize the scenes in which they are present. With that goal in mind, this thesis presents a novel method for the automated detection and sparse 3-D reconstruction of imaged coplanar repetitive patterns. The proposed method applies to a very general class of patterns that encompasses nearly all man-made patterns. Particular contributions include a new set of geometric constraints to eliminate the geometric ambiguity between the imaged and scene pattern, a method to reconstruct the pattern's motif, and a robust framework that successfully detects and reconstructs patterns in the presence of clutter and imaged from lens distorted cameras.

Abstrakt

Opakující se vzory v obrázcích často vznikají díky přítomnosti syntetických, člověkem vytvořených, objektů ve scéně a jsou proto tak poměrně obvyklé v běžných obrázcích. Člověkem vytvořené objekty, mnohdy přítomné ve fotografiích, často obsahují opakující se vzory. Jejich nenáhodně repetitivní struktura porušuje statistické předpoklady, běžně používané v algoritmech počítačového vidění, a negativně tak ovlivňuje jejich výsledky. Tuto repetitivní strukturu lze však detekovat a modelovat a tak využít jejích unikátních vlastností pro charakterizaci scény ve které se nacházejí. Tato práce představuje nový způsob automatické detekce a řídké 3D rekonstrukce ko-planárních repetitivních struktur. Navrhovanou metodu lze využít pro detekci širokého množství tříd repetitivních struktur, které zahrnují skoro všechny syntetické opakující se struktury v obrazech. Významným přínosem této práce je nová sada geometrických pravidel pro eliminaci geometrických nejednoznačností při rekonstrukci, metoda pro rekonstrukci motivu vzoru a robustní systém pro detekci a rekonstrukci vzorů i v případě zákrytů či radiálního zkreslení.

Resumé

Prior to joining Center for Machine Perception (CMP), Mr. Pritts had an established industry career in computer vision and scientific computing. Mr. Pritts was a Lead Engineer for BAE Systems, where he contributed to several US Department of Defense (DARPA) computer-vision research efforts; at NASA, he created gesture-recognition software for remotely controlling robotic arms of the International Space Station; and for Shell Global Solutions, he designed high-performance process-control algorithms.

Mr. Pritts has created novel algorithms for automated camera calibration, object recognition, object classification, and change detection. He received his Bachelor of Science degree in Mathematics from the University of North Texas in 2002.

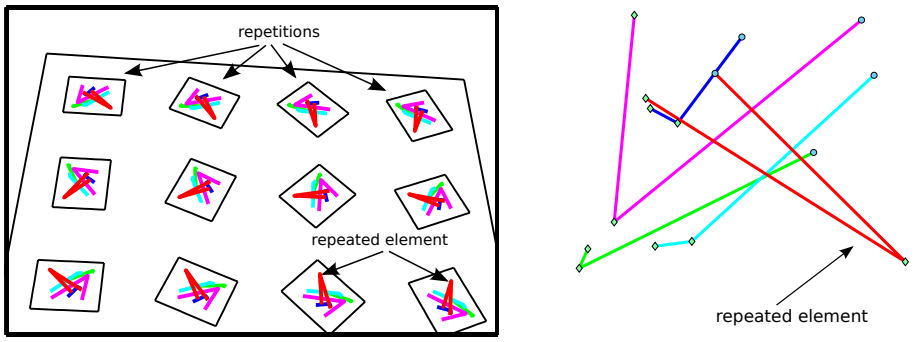
Contents

1 Introduction

Repetitive patterns typically arise from man-made objects and structures and are ubiquitous in image collections. A primary example is urban architecture, which typically contains sets of recurring aesthetic and structural facade elements, a subset of which, *e.g.*, doors and windows, are nearly universally symmetric. The arrangements of repetitions and the symmetry of individual pattern elements are highly discriminative features that can be used to uniquely characterize a scene. Toward that goal, this thesis presents a robust method for the automated detection and sparse reconstruction of imaged co-planar repetitive patterns.

1.1 Pattern terminology

A *repetitive pattern*, or sometimes just *pattern*, is said to be a set of features on a scene plane that can be partitioned into non-singleton, equally-sized subsets called *repetitions*, such that each repetition differs from another by an affinity. Correspondences between repetitions are called *repeated elements*. The motif is an m-tuple of elements, where each motif element corresponds to a repeating element of the pattern. The pattern can be reconstructed through a composition of repetitions of the motif. Note that the dimension of the motif is equal to the cardinality of any repetition. An *imaged repetitive pattern* is a repetitive pattern that is viewed by a perspective camera with lens distortion. The terminology defined above is used for both scene-plane and imaged patterns. Repetitive patterns will be analyzed in their *sparse representation* by local features: the salient elements of the imaged pattern are extracted with a low-level feature detector and are localized by keypoints. The representation puts the pattern in a condensed form for efficient processing while maintaining its distinctiveness.



(a) An imaged repetitive pattern. **(b)** The motif of the pattern.

Figure 1.1 Pattern terminology as shown an in an imaged repetitive pattern in its spare representation by detected features on repeated scene elements.

1.2 Motivation

Two common vision applications that benefit from explicit identification of repetitive patterns are stereo matching and image retrieval. Both applications begin with matching region descriptors, typically Scale-Invariant Features (SIFTs) [16]. In image and particular-object retrieval, SIFTs are quantized, typically by k-means, creating a visual vocabulary where each cluster centroid represents a visual word. Images are represented as bags-of-words, such that word frequency is the tantamount measure for image similarity [30]. The presence of repetitive patterns in images violates the presumption of statistical independence of visual word frequencies [14]. Furthermore, idf weighting does not account for the intra-image feature repetitions presented by patterns. Jegou et al. [14] introduce a set of effective heuristics to account for repeated patterns and textures, but a method to robustly detect repeated patterns would enable a more principled model-based approach.

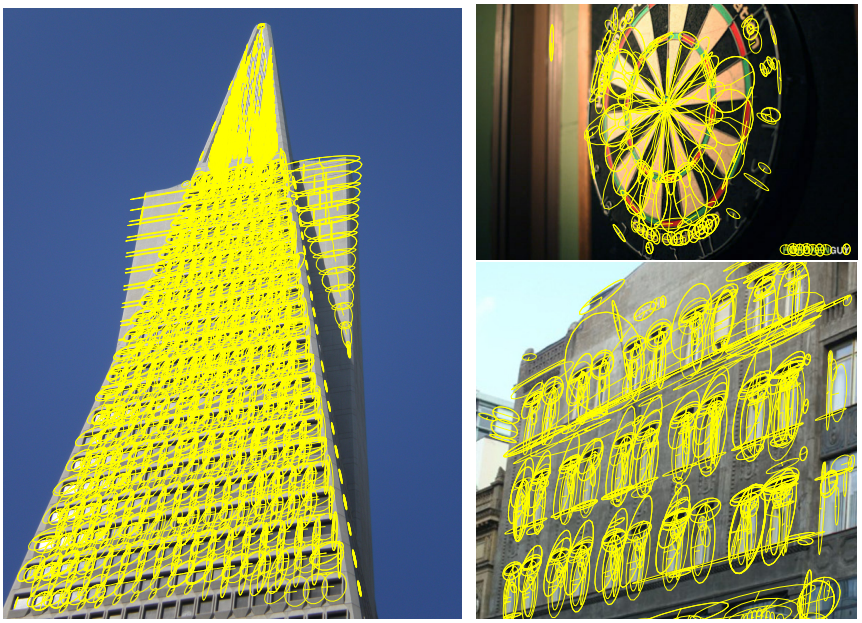
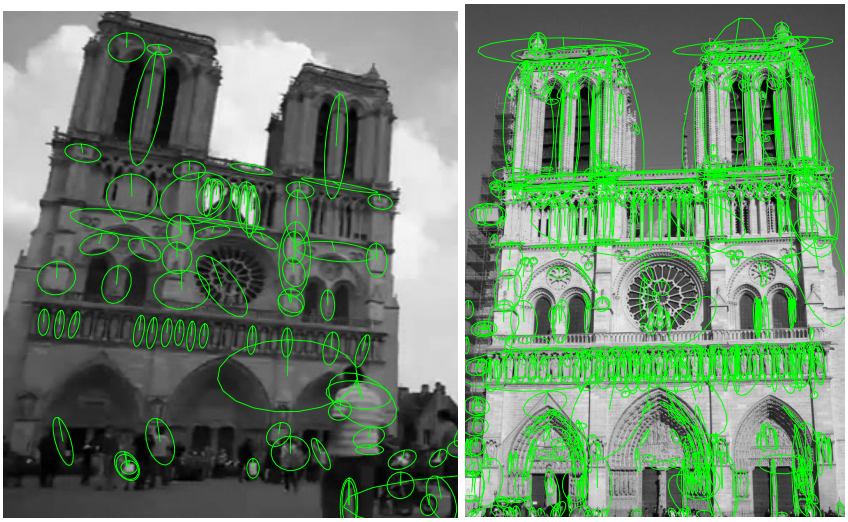


Figure 1.2 "Burstiness examples." Feature repetitions violate independence assumptions of image retrieval algorithms.

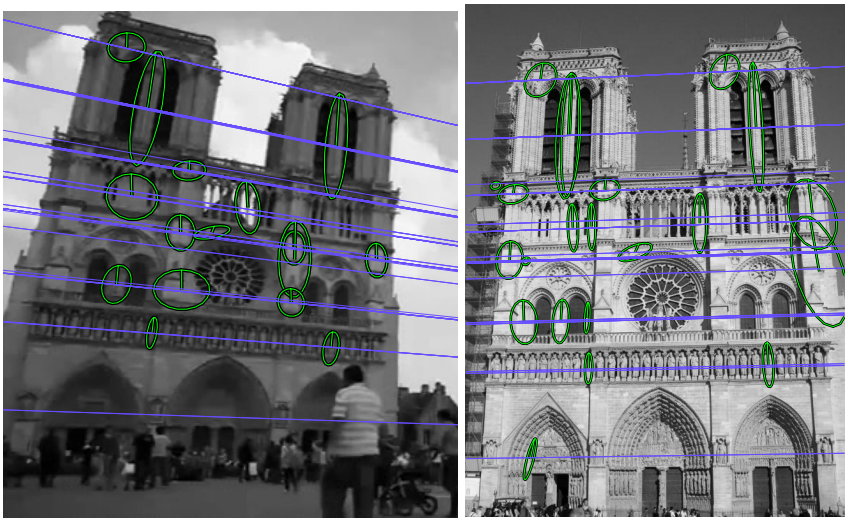
In the case of wide-baseline stereo matching, the presence of a repetitive pattern presents an inter-image matching ambiguity. SIFTs and other common descriptors provide no spatial or semantic context for matching, so matches are made solely based on photometric likenesses. The predictable result is that many mismatches occur between recurring pattern elements. Figure 1.3 demonstrates that in a challenging wide-baseline stereo matching problem, repeated pattern elements either fail to correspond or are erroneously corresponded as inliers to the estimated epipolar geometry.

When matching between patterns with elements that have no unique local texture, the matching problem becomes hopeless without the addition of some global context. Worsening the problem are the effects to matching from viewpoint and lighting changes between the two images, which, despite some designed invariance to ambient effects in local descriptors, can have a detrimental influence on similarity scores. Modeling the repetitive pattern constrains the search for tentative correspondences: geometry and scene semantics can be used to correspond regions that are globally consistent.

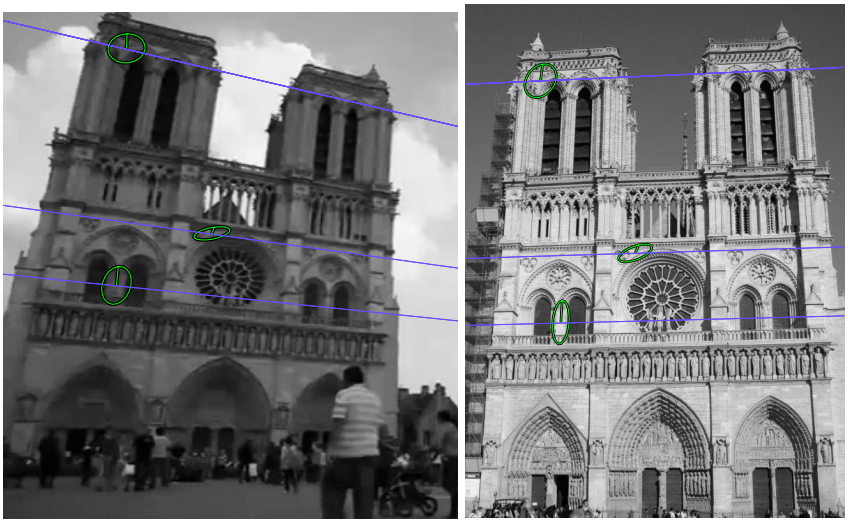
Furthermore, pattern modeling brings benefits mutual to both applications. Scene occlusions can hide large regions of the pattern, reducing the number of available features, which makes



(a) detected MSERs-



(b) correspondences that fit epipolar constraint



(c) correspondences that fit homography constraint

Figure 1.3 Wide-baseline stereo matching problems caused by the presence of a repetitive pattern. Under the balustrade, the statues of the wide horizontal frieze are all detected with MSERs. By chance, some detections correspond and are inliers to the computed epipolar geometry; after spatial verification by homography, none of the detections along the frieze were correctly matched. The consequence is that the accuracy of the estimated two-view geometry suffers.

matching more difficult. Explicitly modeling the pattern enables matching from global context; in other words, hidden regions can be inferred from the model based on the unoccluded regions. Additionally, pattern modeling can be used to increase the quality of local descriptors ex-post. Single view geometry estimated from the pattern can be used to refine the localization of detected features giving a more uniform set of regions of interest. Better localization results in a decrease in the descriptor variance of repeated elements, which has benefits to descriptor quantization and guided sampling. Along similar lines, the multiple views of a repeated element can be fused in a single model, with variations due to lighting and shadows removed.

The advantages of modeling repetitive patterns for common vision tasks are manifold, but reliable fully-automated detection and modeling of repetitive patterns is a challenging problem [34]. Complications arise from patterns that recur at different scales, intervals and orientations, which necessitates a complex model; patterns that include large textureless regions, making it difficult to localize and orient features; and, conversely, patterns that contain dense, highly-textured regions, resulting in a high number of closely grouped features that all look the same.

Pattern detection and modeling can be made more tractable by first observing that in many contexts, the recurring pattern occurs on a scene-plane. Examples include building facades and decorative prints. Most state-of-the-art methods, including the method introduced in this thesis, begin with the planarity assumption.

For planar patterns it is possible to obtain partial camera calibration [3, 12, 28] and to obtain a sparse reconstruction of the pattern up to an affine ambiguity, and, as this thesis will show, contingent on how the pattern recurs, up to a similar ambiguity. This high-level geometric modeling of the pattern provides the geometric and semantic context to resolve the aforementioned complications introduced by the presence of patterns in images for common vision tasks.

1.3 State of the art

Because of their ubiquity in images and importance to multiple problems of computer vision, repetitive patterns continue to receive a lot attention from the research community. Several methods for detecting recurring structure and symmetries have been proposed. In particular, symmetric objects and building facades have received outsized attention in part because the inherent constraints of these two classes of objects make automated pattern detection and modeling more tractable.

Leung and Malik [31] proposed an early solution that constructs a 2-D lattice from an initial patch in a manner analogous to SSD tracking. Schaffalitzky and Zisserman [27] also use local patches to grow a lattice, and assuming coplanarity, explicitly model inter-patch transformations in the warped plane as conjugate translations and rotations. Tuytelaars et al. [32] detect colinear line intersections with a cascaded Hough transform, which are subsequently used in a method that simultaneously detects symmetric fixed structures [33] and affinely rectifies the pattern using projective constraints. Liu et al. [15] model the topological lattice structure of textures using crystallographic group theory. Their approach requires fronto-parallel images. Park et al. [24] formulate pattern detection as a spatial, multi-target tracking problem using mean-shift belief propagation. Doubek et al. [6] develop a shift-invariant descriptor and explicitly consider the pattern during image retrieval and show improved results, but they consider only patterns with translated motifs.

Closer to the development in this thesis, Hong et al. [12] develop constraints based on several classes of symmetries for estimating camera pose and sparse pattern structure. Francios et al. [10] create a virtual mirror of the object and reconstruct it within a multiple-view geometry framework. Their approach requires perfectly symmetric objects, though. Wu et al. [34, 35] also use symmetry to detect translated repetitions in an optimization framework.

1.4 Contributions

This thesis develops an automated non-combinitorial detection and sparse 3-D reconstruction method for imaged co-planar scene patterns with motifs that are transformed intra-pattern by a set of affinities. The method applies to a very general class of patterns which encompass nearly all man-made patterns, including those typically found on building facades, mosaics and decorative prints. Additionally, the presented method works for imaged patterns with significant perspective warping acquired by lens-distorted cameras. In summary, this thesis presents a method that detects and reconstructs patterns in an unconstrained and general image-acquisition setting.

A large subset of contemporary methods use single view geometry to detect and model the repetitive pattern [12, 27, 32, 34, 35]. But these techniques impose several requirements on scene content that restrict the class of imaged patterns to be considered. Common to many of the cited methods is the need for vanishing lines or the presence of certain types of symmetries. To the author's knowledge, no technique has addressed patterns imaged with significant lens distortion.

The proposed approach exclusively uses repeated pattern elements for single-view geometry estimation. The approach greatly expands the class of images in which patterns can be detected because the need to detect extra-pattern features, such as vanishing lines, is obviated. Furthermore, the only requirements of the method is that the repeated elements can be mapped to each other via an affinity in the scene plane. This is a very general assumption that covers all commonly seen man-made repetitive patterns.

1.5 Thesis structure

3-D sparse pattern reconstruction begins with the extraction of low-level local features and successively upgrades the pattern's representation by removing geometric ambiguities from the imaged pattern and putting a global context to the local features. Analogously, this thesis will follow that development. Chapter 2 provides background and context to local-feature based representations of scenes. A state-of-the-art feature representation is reviewed and its adoption is justified. Notations and conventions are introduced for working with low-level features that will be useful in the development of novel single-view geometry. Chapter 3 begins the process of inferring pattern configuration from local features. A clustering method based on local texture similarity is introduced to establish inter-repetition pattern correspondence. Chapter 4 develops a single-view geometry pipeline that uses several geometric constraints to reduce geometric ambiguity between the imaged pattern and scene pattern. Putting the pattern in a rectified plane enables the comparison of relative angles and distances. Spatial clustering is used to estimate an initial guess at the pattern's motif and repetition configuration. Chapter 5 puts the estimators developed in Chapter 4 into a robust framework and generalizes the approach to work for cameras with significant radial lens distortion. A nonlinear estimator is developed that gives an optimal reconstruction of the pattern. Experiments are presented that demonstrate the stability and accuracy of the framework.

2 Image Features

Digital images are acquired by recording the radiance from a sampled set of light rays projected onto a CCD or CMOS sensor. Projection destroys the 3-D structure of the scene and what remains are the samples, which retain little of the physical meaning of the real-world objects that they represent. Scene understanding begins at the lowest level: from a 2-D array of radiance values, *i.e.* pixels, which is the output of the camera. While some scene understanding frameworks use pixel values directly, it is often beneficial, both in terms of run-time performance and algorithm efficacy, to use derived image features such as corners, contours, blobs and lines that capture salient quantities of the scene. In particular, regions with high luminance gradient magnitude arise from high-contrast textures and 3-D depth discontinuities, thus capturing structural attributes of the underlying scene. Furthermore, there is an excellent precedent for starting with higher-level quantities as there are direct analogs to many common feature detectors typically used in computer vision with mid-level brain function in biological vision [1].

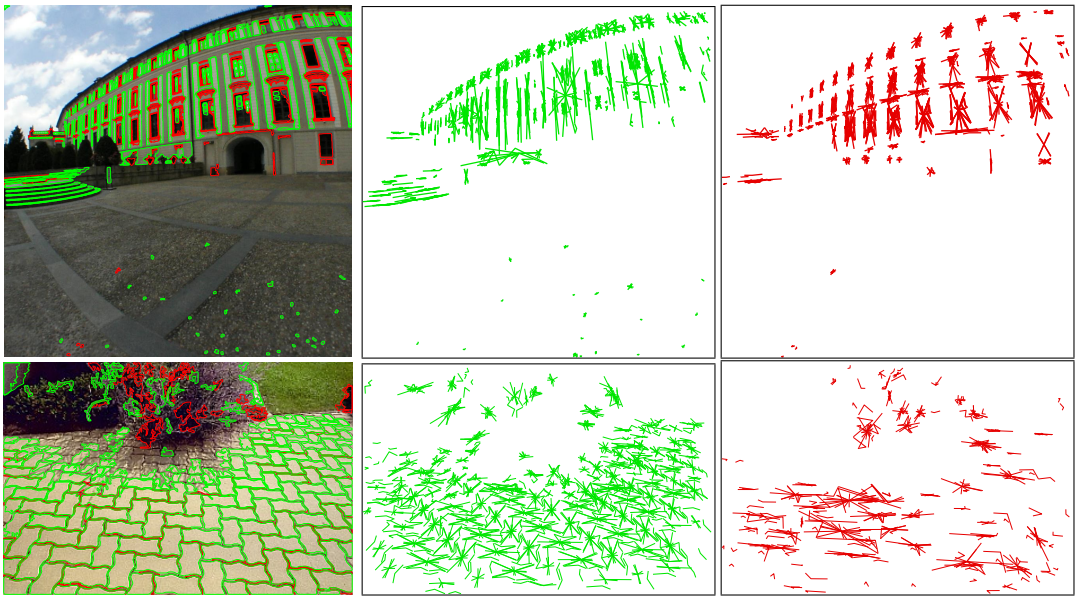
If a repetitive pattern is present in the image, it typically has a higher level of structure and texture than the surrounding clutter. This fact suggests that high-level feature extraction is a good way to immediately discard large image regions of low-texture that are unlikely to be part of a pattern. The extracted features are characterized such that they can be compared to find features that are similar. Features that represent repeated pattern elements should cluster, a fact that will be used to make a tentative first guess at the pattern's structure.

Critical to the reliability of vision applications is the repeatability of detections under changes of illumination, camera position and orientation, and other ambient imaging conditions [21]. Repeatability is a particularly important property for the task of pattern detection and reconstruction since it is the detections of repetitions of particular pattern elements that will enable the estimation of the pattern's motif. The feature detection pipeline presented in this chapter was assembled to detect 2-D structural elements in a repeatable way. Significant consideration is also given to the concise representation of image structure for the design of efficient feature matching and single-view geometry algorithms.

2.1 Detection

The importance of contours for scene understanding is supported by a recent study [4, 7] which found that generally subjects are able to judge 3D surface normals of an object depicted by a line-drawing almost as accurately as for objects depicted by a shaded image. Luminance level sets often correspond to object contours; consequently, there has been a lot of effort in the computer-vision research community to design robust level-set detectors [21]. In particular, Matas et al. [18, 19, 22, 25] develop a highly effective pipeline for robustly detecting and compactly representing level sets. Level sets are detected with the Maximally Stable Extremal Region detector [18]. A *Maximally Stable Extremal Region* (MSER) is a connected component of a luminance image where all member pixels are brighter (MSER+) or darker (MSER-) than all surrounding pixels. The "maximally stable" property denotes that the region boundary remains stable over a user-supplied gap of luminance thresholds. MSERs are highly robust to both geometric and photometric changes: they are covariant to diffeomorphisms, enabling matching over diverse camera geometries; and they are invariant to monotone changes of luminance, which can be caused by ambient scene effects or gamma correction. Affirming

these theoretical advantages, Mikolajczyk et al. [21] demonstrated the high-repeatability of the MSER detector on challenging data sets.



(a) MSER boundaries **(b)** LAFs from MSER+ **(c)** LAFs from MSER-

Figure 2.1 MSERs are detected, from which LAFs are derived. Note that the salient structure is retained by the set of LAF constructions, and the repetitive pattern is easily discernible in its sparse representation.

Dealing directly with MSER boundaries is data-intensive and does not directly allow the use of point-based single-view geometry methods. Matas et al. [19] introduce *local affine frames* (LAFs) to compactly represent the most salient differential geometric properties of the MSER while preserving robustness. Affine coordinate systems are constructed from 3-tuples of distinctive affine-covariant points on the boundary and interior of the MSER, such as curvature extrema, concavities and convexities, bi-tangents, inflection points and moments of the MSER [23].

Note that during LAF construction, some geometric covariance of the MSER is lost. However, as shown in Section 4.2.1, the perspective distortion of an image region local to a pattern element can be approximately modeled by an affinity. This property suggests that the use of affine-covariant features will be sufficient to detect repeated pattern elements, and a more convenient representation can be used in lieu of directly using the MSER boundaries.

2.2 Description

Outputted from LAF construction are affine coordinate systems constructed from triplets of affine-covariant points on the MSER that exhibit distinctive differential geometric properties. Naturally, these affine frames define *distinguished regions* (DRs), since they are well-located and structured. Typically, the distinguished region is extended beyond its defining affine frame to capture some of the surrounding area as well. To establish a tentative spatial structure for the pattern, matching of similar distinguished regions in the image must be established. Again, following Matas et al. , the distinguished regions are normalized via their local coordinate systems to an orthonormal basis (Matas et al. refer to the orthonormal basis as the *canonical frame*). Locally, projective distortions can be modeled by affine transformations (see Section 4.2.1), so

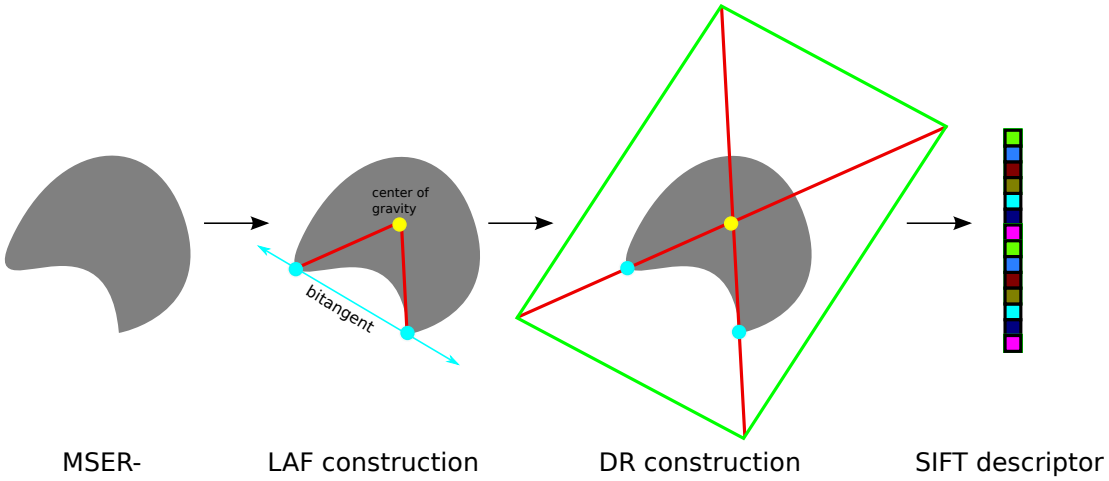


Figure 2.2 Local feature representations. A level set is detected in the image as an MSER-; salient differential geometric points are detected on the MSER- from which a LAF is constructed; the DR captures surrounding texture of the LAF; the texture of the DR is represented as a SIFT. LAFs and SIFTs are kept as persistent tokens for further processing.

distinguished regions that contain the same repeated scene element of a pattern should appear alike after normalization.

2.2.1 Scale Invariant Feature Transform (SIFT)

A representation for DRs is needed so that they can be matched to establish repetition correspondence. Early attempts directly compared pixels of the DR [19] via normalized cross-correlation. Despite geometric and photometric normalization, small perturbations in this representation were found to have an out-sized impact. Patch Representation by a histogram of gradients [5, 16] was found to greatly enhance robustness to small changes in the normalization. The Scale Invariant Feature Transform (SIFT) [16] is one of the most widely used gradient-based methods for DR representation. The descriptors are referred to as SIFTs, and their robustness has been affirmed in many use cases, including in a survey of state-of-the-art region descriptors by Mikolajczyk et al. [20].

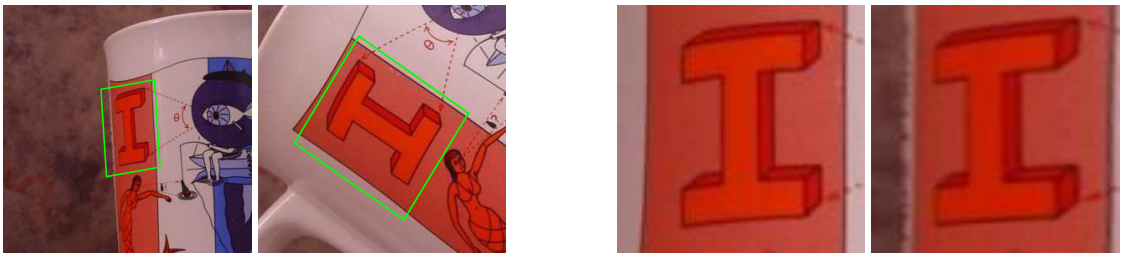
The SIFT descriptor is a 128-D histogram of quantized gradient values. A similarity measure on SIFTs is needed to establish tentative correspondences. A recent measure that has shown improved matching performance over cosine similarity and and Euclidean metric is RootSIFT, introduced by Arandjelovic et al. [26]. The key insight to RootSIFT is that while SIFTs were originally designed for use with the Euclidean metric, they are histograms, and would likely benefit from the use of a histogram similarity measure. Arandjelovic et al. choose the Hellinger kernel and show a significant performance increase across several common computer vision tasks. The Hellinger kernel is defined as

$$H(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \sqrt{x_i y_i},$$

and can be computed with euclidean distance by a clever transformation: (i) $L1$ normalize the SIFT vector (originally it has unit $L2$ norm); (ii) square root each element. With this transformation, it can be shown that

$$\|\sqrt{\mathbf{x}} - \sqrt{\mathbf{y}}\|_2^2 = 2 - 2H(\mathbf{x}, \mathbf{y}).$$

The implication is that with a simple transformation, RootSIFT can be used in algorithms that require Lp-norm distances, *e.g.*, k-means [26], which will prove critical for the SIFT clustering techniques used to establish pattern repetition correspondence introduced in Chapter 3.



(a) Distinguished regions (DRs) **(b)** appearance in canonical frame

Figure 2.3 LAF constructions locate regions of high texture, called distinguished regions (green boxes). DRs are normalized and oriented to the canonical frame where they look alike. Images provided by Stepan Obdržálek.

2.3 Sparse representation of the imaged pattern

Pattern reconstruction will make use of both the photometric information provided by SIFTs, and the spatial context provided by their corresponding LAFs. SIFTs will be used to establish a tentative correspondence between repeated pattern elements. LAFs will be used exclusively as the fundamental tokens for the image features used to sparsely represent the imaged pattern. Ultimately, the 3-D reconstruction of the pattern motif will be estimated from the rectified 3-tuples of LAF points. LAFs are highly appropriate for representing the structure of imaged repetitive patterns because they encode rich differential geometric properties, derive high repeatability from the underlying MSERs, and because their affine-covariance allows matching of repeated scene elements across disparate regions of the image. In the following development, it will be shown how the high discriminability of SIFTs and LAFs can be used to cluster features of repeated elements and infer pattern geometry.

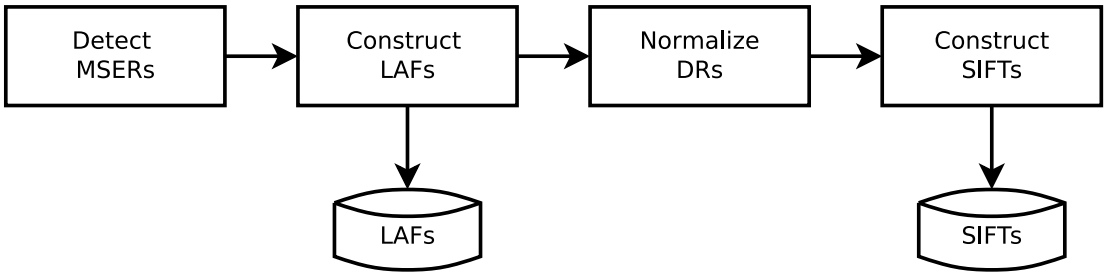


Figure 2.4 Feature extraction pipeline. LAFs are derived from MSERs and are used as the fundamental image token for the sparse representation of the repetitive pattern. LAFs define distinguished regions (*i.e.* DRs are LAFs plus some surrounding area; they are not persistent tokens in the pipeline), which are normalized to the canonical frame. SIFTs are generated from the normalized frames which gives a compact photometric description of the corresponding LAF.

2.4 Working with LAFs

In the subsequent chapters, LAFs will be the image measurement that will be used to spatially verify tentative correspondences, reduce geometric ambiguity between the imaged pattern and

scene pattern, and to estimate the motif. Thus, a compact notation is required for the upcoming development. Recall from Section 2.1 that a LAF is an ordered 3-tuple of affine covariant points. The primary representation of a LAF will be in homogeneous coordinates,

$$L = (\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3) \quad \mathbf{u}_i = \begin{pmatrix} x_i & y_i & 1 \end{pmatrix}^\top.$$

The *LAF origin* is always $\mathbf{u}_2$ and is denoted as the yellow point of the LAF construction in Figure 2.2. LAF points are ordered such that they form a right-handed coordinate system with the positive z-axis coincident with the principal ray of the camera. The set of LAF constructions derived from MSERs is denoted

$$\text{LAFS} = \{ L^{(1)}, L^{(2)}, \ldots, L^{(m)} \}$$

It will occasionally be necessary to reference the individual components of the LAF. For example, the notation to reference the origin of a particular LAF is $\mathbf{u}_2^{(j)}$, and analogously, to reference the inhomogeneous coordinates of the origin, $\left(x_2^{(j)}, y_2^{(j)}\right)$. *LAF canonicalization* is the transformation of the affine coordinate system and surrounding LAFs to the orthonormal basis. It is a useful operation because affine ambiguities are removed between tentatively corresponding LAFs and their surrounding regions, which allows distances and angles to be directly compared during clustering and spatial verification. The notation for a canonical LAF is

$$L_\llcorner = (\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3) \qquad \text{where } \mathbf{n}_1 = \begin{pmatrix} 1 & 0 & 1 \end{pmatrix}, \mathbf{n}_2 = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \mathbf{n}_2 = \begin{pmatrix} 0 & 1 & 1 \end{pmatrix}.$$

Let $\mathtt{N}^{(j)}$ to be the affine transform (a change of basis) that takes LAF $L^{(j)}$ to the canonical frame, denoted as

$$L_\llcorner = \mathtt{N}^{(j)} L^{(j)}.$$

Another LAF property that will prove useful in Chapter 4 is LAF scale. LAF scale is the image area covered by the parralelogram defined by the affine covariant points of the LAF (see Figure 2.2 to see how the parallelogram is constructed). The laf scale is given by

$$\text{scale}(L^{(j)}) := \det\left(\mathtt{N}^{(\mathtt{j})^{-1}}\right). \tag{2.1}$$

3 Tentative pattern structure from SIFT clustering

Recall from Chapter 2 that the persistent tokens emitted from the image pipeline are LAFs and SIFTs. A LAF defines a 3-tuple of distinctive affine-covariant points in the image, and its corresponding SIFT characterizes the texture within the LAF and surrounding neighborhood with an affine invariant descriptor (the affine invarinace arises from the LAF construction). These are local quantities and there exists no correspondence between LAFs; *i.e.*, there is no notion that LAFs might represent the same repeated scene element of an imaged pattern. This chapter proposes a clustering method to establish a tentative correspondence between LAFs that represent the same repeated element. The clustering approach creates connected components of like-textured SIFTs (inducing the same connectedness on the corresponding LAFs). This initial tentative correspondence is used to build the structure of the pattern in subsequent steps.

3.1 Finding repeats from SIFTs

The SIFTs provide a way to establish a tentative correspondence between LAFs which coincide with repeated pattern elements. A SIFT descriptor is a 128-D vector that effectively characterizes the texture of a region around a keypoint [16]. The approach in this thesis defines the region as a scaled area around a LAF construction, also called a distinguished region (DR). A similarity measure is chosen so that SIFT likeness can be measured. The most common measures used for SIFTs are the euclidean metric, cosine similarity and earth mover's distance. In stereo matching, a common approach is to establish tentative correspondences between SIFTs if the ratio of distance from the closest neighbor to the distance of the second closest is less than some threshold, usually around 0.8 [16]. In the image retrieval context, k-means is commonly used to cluster like SIFTs and the centroids of the clusters are used for visual words, which are used in aggregate to characterize the image [30]. Unfortunately, neither of these matching approaches is applicable to establishing tentative correspondences between repeated pattern elements. The ratio measure of Lowe is quickly dismissed; in fact, the correspondences of interest (*i.e.*, the repeated elements) will not distinguish themselves from the k-th closest match because all imaged repetition will look very similar. k-means clustering is not appropriate because there is no way to know the number of repeated elements in the pattern apriori. For large image databases, the performance of the retrieval engine is not so sensitive to small changes in the number of clusters. But in the context of repetitive pattern detection, there is only sparse data extracted from one image and the number of clusters is small, so small perturbations have a larger impact. The conclusion is that a different approach is needed.

3.2 Establishing repetition correspondence

An efficient method that can partition the SIFTs into clusters that represent pattern element repetitions without an apriori guess on the number of repeated pattern elements is sought. *Spectral clustering* is an efficient clustering method that meets the above requirements and has several fundamental advantages over traditional algorithms like k-means and agglomerative clustering. An in-depth exposition on spectral clustering is given by Luxburg [17].

3.2.1 Similarity graphs

A *similarity graph* is a weighted adjacency matrix $W = [w_{ij}]$ such that the vertices (row and column indices) are data points (in our case SIFTs) and the edges (matrix entries) are the similarity measures of the data points. With a graph representation of SIFT similarity, SIFT clustering can be cast as a spectral graph theory problem: find a graph partition so that inter-cluster edge weights are low (implying dissimilarity of cluster members), and intra-cluster edge weights are high (implying high similarity of cluster members). The solution to this problem is addressed by spectral clustering.

Euclidean distance of the RootSIFT transformed SIFT descriptors (see Section 2.2.1 for details) is used to measure similarity,

$$s_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|_2, \qquad \mathbf{s}_i, \mathbf{s}_j \text{ are two SIFTs.}$$

A similarity graph $G = (V, E)$ is constructed where the vertices v_i represent SIFT descriptors and the edges are weighted so that

$$w_{ij} = \begin{cases} e^{\frac{-s_{ij}^2}{2\sigma^2}} & \text{if } s_{ij} < t \qquad t \text{ a loose matching threshold} \\ 0 & \text{otherwise.} \end{cases}$$

The parameter σ is user-supplied and is set experimentally from fitting distributions to SIFT similarity scores on an image corpus. For the exponential weighting used, it represents the standard deviation of similarity scores assuming that they are normally distributed. The threshold parameter t is chosen so that it is very unlikely that two SIFTs with a similarity score exceeding t will match. Even with t chosen conservatively, the hard threshold creates a lot of sparsity in the similarity graph immediately enabling the graph to be partitioned into a set of connected components,

$$\mathcal{W} = \{W_i \mid W_i \text{ a connected submatrix of } W\}$$

Splitting the graph into components before the eigen-decomposition of the normlalized Laplacian of the weight matrix (see Algorithm 1) reduces run time; note that eigen-decomposition L is $\mathrm{O}\big(|V|^3\big)$ versus finding connected components which is $\mathrm{O}\big(|V|\big)$.

3.2.2 Clustering by spectral analysis

For the weight matrix W_i of each connected component, normalized spectral clustering as detailed in Algorithm 1 is performed. The maximum eigen-gap of a fraction of the largest generalized eigenvalues is found to automatically determine the number of clusters; *e.g.*, if the maximum eigen-gap occurs at eigenvalue λ_k, the number of clusters is set to k. The SIFTs are projected to the eigenspace and clustered there by k-means with the number of clusters determined by the eigengap analysis. The results of k-means is a clustering of the SIFTs and their corresponding LAFs. Clusters are aggregated over all W_i into a set,

$$\mathcal{C} = \{\, C_1, C_2, \ldots, C_k \,\}$$

The set $\mathcal{C}$ of clusters represents a tentative guess at correspondence between repeated elements of the repetitive pattern.

3.2.3 Discarding bad clusters

Spectral clustering by SIFT distance gives a tentative guess at correspondence between imaged repeated elements. While the method successfully identifies "good" clusters, *i.e.* clusters with

Algorithm 1 Normalized spectral clustering

procedure CLUSTERSIFTS($W = [w_{ij}]_{n\times n}$) ▷ Input is weighted adjacency matrix
  $D \leftarrow [d_{ij}]$ where $d_{ij} = \begin{cases} \sum_j w_{ij} & i = j \\ 0 & \text{otherwise.} \end{cases}$
  $L \leftarrow I - D^{-1/2} W D^{1/2}$ ▷ construct normalized Laplacian
  Solve $Lu = \lambda V u \quad \forall i \in 1 \ldots n/4$ ▷ generalized eigenvalue problem
  $k \leftarrow \arg\max_k |\lambda_k - \lambda_{k+1}|$ ▷ eigen-gap determines the number of clusters
  $U \leftarrow [u_1 u_2 \ldots u_k]$
  $y_i \in \mathbb{R}^k \leftarrow i$th row of U
  $C_1, \ldots, C_k \leftarrow$ KMEANS$((y_i)_{i=1,\ldots,n})$ ▷ cluster transformed SIFTs
  return $C_1, \ldots, C_k$
end procedure

a majority of good correspondences, there are problems with over-segmentation, inclusion of bad correspondences, and clustering of high frequency textures (*e.g.* small corner like features throughout the image) that are not repeated elements. The first two problems will be addressed when spatial information is considered. The last problem will be addressed by examining the statistics of SIFT distances.

Burghouts et al. [2] show theoretically and empirically that L_p-norms from one descriptor vector (*e.g.* SIFT) to other vectors are Weibull-distributed if the descriptor values are correlated and non-identically distributed. This fits the use case in the above development and gives some hope that SIFT match scores can be partitioned into match and non-match classes via unsupervised learning. This approach was attempted and failed. The key insight is that the match score distributions are conditioned on each feature: *i.e.*, the feature's structure and position in the image greatly influence its matching scores. Even softening this requirement to conditioning distributions on cluster membership does not help much. In one image, there is just not enough descriptors per cluster to estimate Weibull distributions.

In lieu of a full estimate of the match score distribution, the median absolute deviation (MAD) of a cluster is tested. MAD is a *robust* measure of statistical dispersion [13]; in this context it estimates the variability of intra-cluster pairwise SIFT distances d_i,

$$\text{MAD}(C_k) = \text{median}_i(|d_i - \text{median}_j(d_j)|)$$

The expectation is that the variability in clusters of repeated pattern elements will be lower than that of clusters of smaller high frequency features. Some "good" clusters will have bad matches, but MAD is robust to small corruptions, so a good dispersion estimate will still be obtained for the matching SIFTs. A conservative threshold was experimentally determined, and clusters whose MAD estimate exceeds the threshold are discarded,

$$\text{class}(C_k) = \begin{cases} \text{“}good\text{''} & \text{if MAD}(C_k) < t \\ \text{“}bad\text{''} & \text{otherwise} \end{cases}$$

3.3 LAF clusters

The result of SIFT clustering is the collection of clusters $\mathcal{C}$, such that each cluster $C \in \mathcal{C}$ is a set of tentative correspondences between repeated elements of the pattern. SIFTs correspond one-to-one with the LAFs, which define the distinguished region from which the SIFT descriptors are generated. So, of course, a clustering of the LAFs is achieved by $\mathcal{C}$, too. Some additional notation for LAF clusters will be of use. A LAF cluster C_i is an m-tuple of intergers indexing

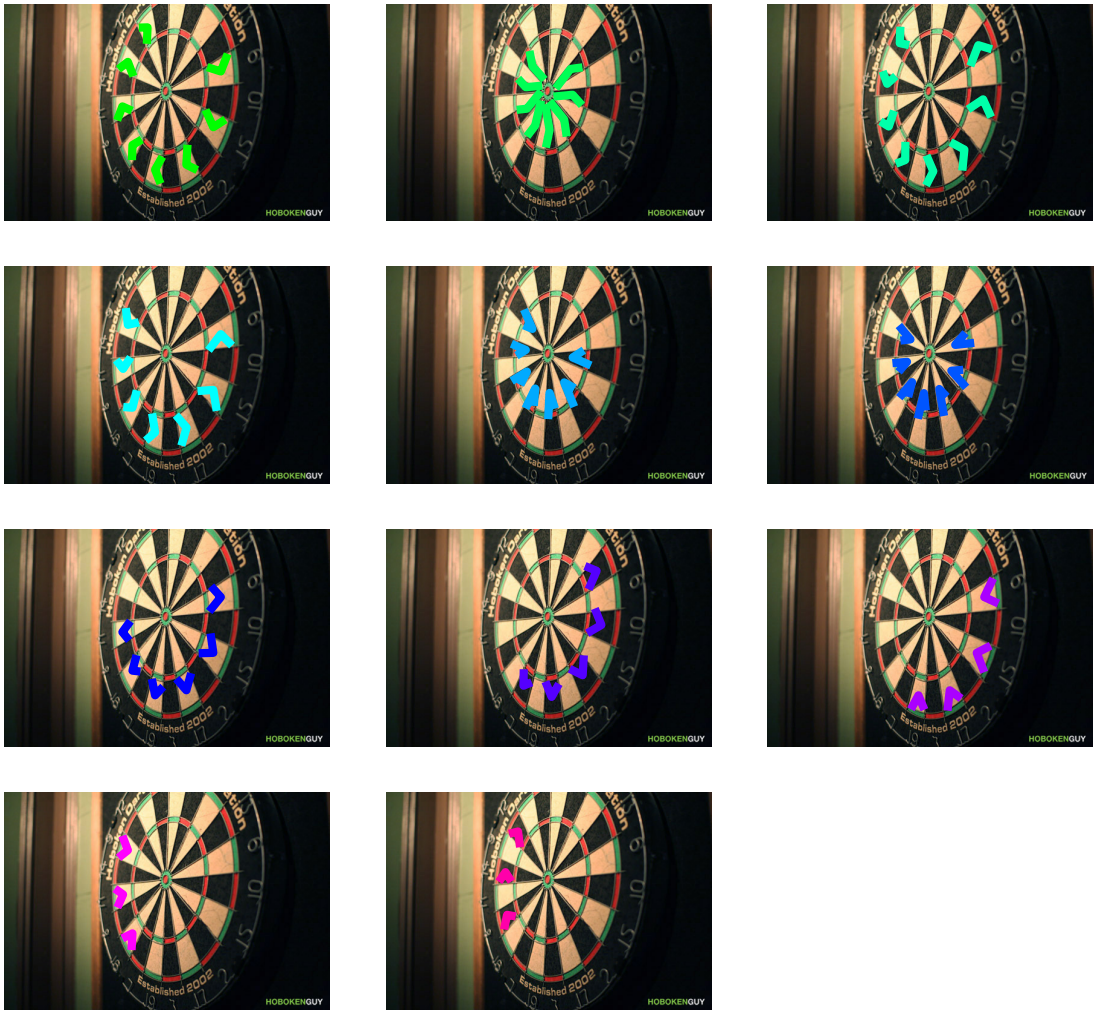


Figure 3.1 Dartboard, result from automated clustering. 11 clusters automatically detected through eigengap analysis. No bad clusters returned, but there are some missed matches, particularly in rows 3 and 4.

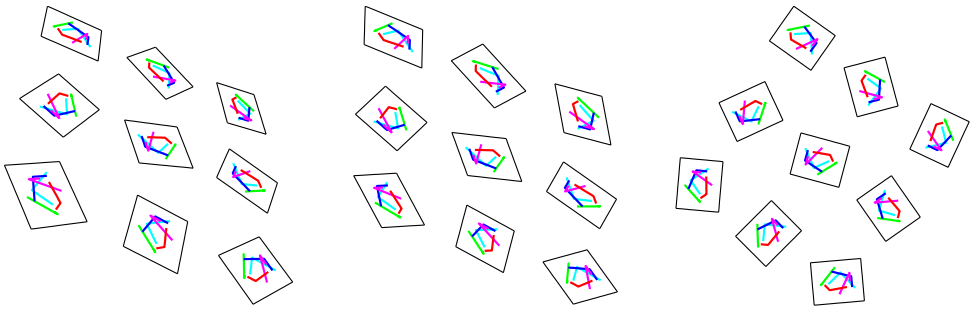
into the set of detected LAFs such that all indexed LAFs are tentatively corresponded with each other. For example, the jth LAF in the ith clusters is denoted $L^{C_i(j)}$ (see Section 2.4 for LAF denotations). In Chapter 4, the spatial context provided by the LAFs will be used for eliminating mismatches from the tentative correpsondences, rectifying the imaged pattern, and for estimating the motif.

Figure 3.2 outdoor scene, result from automated clustering. 12 clusters automatically detected.

4 A linear method for sparse 3-D repetitive-pattern reconstruction

This chapter develops the single-view geometry necessary for the linear sparse 3-D reconstruction of a co-planar repetitive pattern. The method presented works on virtually all man-made patterns. It requires as few as two repeated features, *i.e.*, four LAFs, making it appropriate for use in hypothesis and test frameworks, such as RANSAC [8]. Synthetic experiments show that the proposed methods are robust to feature detection noise and are stable across a very broad range of scene geometries. Reconstructions on real-world imagery affirm the algorithm's effectiveness.

The method proceeds by successively reducing the geometric ambiguity between the scene and imaged pattern, where, at each geometric upgrade, additional constraints on repeated pattern elements are introduced. The results of the linear pipeline are an estimation of the rectified pattern's motif, and the set of transformations needed to orient repetitions of the motif, which can be used for the rectified 3-D sparse reconstruction of the pattern.



(a) perpsective ambiguity **(b)** affine ambiguity **(c)** similarity ambiguity

Figure 4.1 The linear pipeline reduces geometric ambiguity between the imaged plane and scene plane through a series of linear estimations. If a repeitition is rotated, the ambiguity can be reduced to a similarity as in (c).

4.1 Overview

Recall from Section 3.3 that LAF clustering gives a tentative inter-repetition feature correspondence; *i.e.*, if there is a repetitive pattern in the image, then some subset of the LAF clusters will represent repeated pattern elements. In the scene plane, repeated elements have the same scale, as opposed to the their imaged counterparts, *i.e.*, LAFs, which are detected at different scales because of perspective camera geometry. This relative scale discrepancy between imaged repeated elements and their scene-plane counterparts can be used to rectify the imaged scene plane [3].

After rectification, repetitions differ pair-wise by at most an affinity. Each repetition is transformed to the canonical frame, which enables direct comparison of distance and angle between repeated elements. The tentative correspondence between repeated elements is improved with spatial verification and guided sampling [11]. Subsequently, the pattern is partitioned into its constituent repetitions from which a rectified pattern motif is estimated. If any repetition is a

rotation or reflection of the motif, the geometric ambiguity between the scene plane and the imaged pattern can be reduced to a similarity.

The results from the linear methods proposed, in aggregate, are: scene-plane rectification; partial camera calibration, most importantly, the relative orientation of the camera with respect to the scene plane; sparse motif reconstruction; the relative orientations of the repetitions; and a sparse 3-D reconstruction of the pattern.

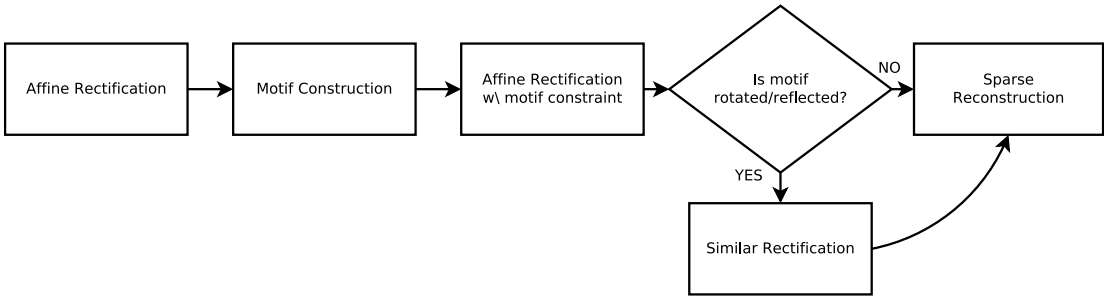


Figure 4.2 Linear pipeline for sparse 3-D reconstruction

4.2 Pattern rectification from scale change of LAFs

An algebraic constraint on the scale change of a scene-plane feature transformed by a homography was introduced by Chum et al. in [3]. This scale constraint directly leads to constraints on the projective entries of the transforming homography and on the position-independent scale change between matching affine-rectified features and their scene-plane counterparts. The constraints can be stacked in a design matrix to give an accurate linear method for estimating the rectifying homography $\mathtt{H}_\infty$ of the scene-plane containing the repetitive pattern.

4.2.1 Local Scale Change in an Image

The relative scale change of a feature from its scene-plane area to its imaged area can only be measured in the Euclidean plane, so the projective transformation of $\mathbf{u} = (\zeta x, \zeta y, \zeta)$,

$$\mathtt{H}\mathbf{u} = \begin{pmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & \lambda \end{pmatrix} \mathbf{u} = \begin{pmatrix} \mathbf{h}_1^\top & h_3 \\ \mathbf{h}_2^\top & h_6 \\ \mathbf{h}_3^\top & \lambda \end{pmatrix} \mathbf{u}, \tag{4.1}$$

will be represented in its inhomogeneous form by setting $\lambda = 1$ and $\zeta = 1$. Denoting the inhomogeneous representation of $\mathbf{u}$ as $\boldsymbol{u} = (x, y)$, the image of the transformed point is the vector-valued function

$$\boldsymbol{u}'_{\mathtt{H}}(\boldsymbol{u}) = \frac{\begin{pmatrix} \mathbf{h}_1^\top \\ \mathbf{h}_2^\top \end{pmatrix} \boldsymbol{u} - \begin{pmatrix} h_3 \\ h_6 \end{pmatrix}}{\mathbf{h}_3^\top \boldsymbol{u} + 1}. \tag{4.2}$$

Fixing $\lambda = 1$ is incorrect if the image of the origin is an infinite point of $\mathcal{P}^2$. Changing the coordinate system by an affinity that moves the image of the origin away from any infinite points corrects the geometry. Note that the choice of affinity is dependent on the configuration of the detected features. Linearizing about the point $\boldsymbol{u}$ with the first-order Taylor expansion gives

$$\boldsymbol{u}'_{\mathtt{H}}(\boldsymbol{u} + \boldsymbol{\delta}_{\boldsymbol{u}}) = \boldsymbol{u}'_{\mathtt{H}}(\boldsymbol{u}) + \nabla \boldsymbol{u}'_{\mathtt{H}}(\boldsymbol{u}) \boldsymbol{\delta}_{\boldsymbol{u}}, \tag{4.3}$$

where $\boldsymbol{\delta_u}$ is a local displacement. Now consider the first-order term, which is responsible for the change of scale,

$$\nabla \boldsymbol{u}'_{\mathtt{H}}(\boldsymbol{u}) = \frac{\left(\mathbf{h}_3^\top \boldsymbol{u} + 1\right) \begin{pmatrix} \mathbf{h}_1^\top \\ \mathbf{h}_2^\top \end{pmatrix} - \begin{pmatrix} \mathbf{h}_1^\top \\ \mathbf{h}_2^\top \end{pmatrix} \boldsymbol{u}\mathbf{h}_3^\top + \begin{pmatrix} h_3 \\ h_6 \end{pmatrix} \mathbf{h}_3^\top}{(\mathbf{h}_3^\top \boldsymbol{u} + 1)^2}. \tag{4.4}$$

The Jacobian $\nabla \boldsymbol{u}'_{\mathtt{H}}(\boldsymbol{u})$ is a 2×2 affinity and thus local change of scale is approximated by the determinant of the Jacobian,

$$\det\left(\nabla \boldsymbol{u}'_{\mathtt{H}}(\boldsymbol{u})\right). \tag{4.5}$$

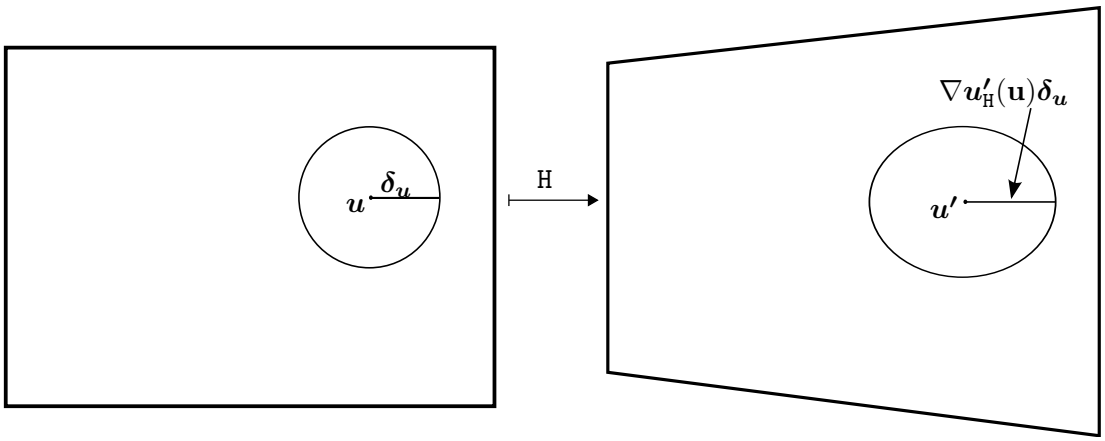


Figure 4.3 The approximate scale change $\nabla \boldsymbol{u}'_{\mathtt{H}}(\mathbf{u})$ of a region defined by point $\boldsymbol{u}$ and radius $\boldsymbol{\delta_u}$ after transformation by $\mathtt{H}$. The relative scale change is given by $\det \nabla \boldsymbol{u}'_{\mathtt{H}}(\boldsymbol{u})$.

4.2.2 A linear constraint on local scale change

As before, let $\mathbf{u} = (\zeta x, \zeta y, \zeta)$ and $\boldsymbol{u} = (x, y)$, and define the scale change from homography $\mathtt{H}$ as

$$s(\mathtt{H}, \mathbf{u}) \equiv \det\left(\nabla \boldsymbol{u}'_{\mathtt{H}}(\boldsymbol{u})\right).$$

By fixing $\lambda = 1$, a homography of the form in equation 4.1 has the following decomposition

$$\underbrace{\begin{pmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & 1 \end{pmatrix}}_{\mathtt{H}} = \underbrace{\begin{pmatrix} h_1 - h_3h_7 & h_2 - h_3h_8 & h_3 \\ h_4 - h_6h_7 & h_5 - h_6h_8 & h_6 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathtt{A}} \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ h_7 & h_8 & 0 \end{pmatrix}}_{\widetilde{\mathtt{H}}}, \tag{4.6}$$

so the scale change of $\mathtt{H}$ can be expressed in terms of its decomposition $\mathtt{H} = \mathtt{A}\,\widetilde{\mathtt{H}}$ as

$$s(\mathtt{H}, \mathbf{u}) = \det\left(\nabla \underset{\sim}{\boldsymbol{u}}'_{\mathtt{AH}}(\mathbf{u})\right) = \det(\mathtt{A}) \det\left(\nabla \underset{\sim}{\boldsymbol{u}}'_{\mathtt{H}}(\mathbf{u})\right). \tag{4.7}$$

Note that $s(\mathtt{H}, \mathbf{u})$ is the product of two affinities. There are two advantages to decomposition 4.7: the scale changes due to the affine and projective components of the homography are separated, and the scale change due to the affine component is global, *i.e.*, not position dependent. Furthermore, we have an expression for the local scale change $\nabla \boldsymbol{u}'$, namely equation 4.4,

which can be simplified by considering the special structure of $\widetilde{\mathtt{H}}$. Note that for $\widetilde{\mathtt{H}}$, $\begin{pmatrix}\widetilde{\mathbf{h}}_1^\top \\ \widetilde{\mathbf{h}}_2^\top\end{pmatrix}$ is I_2 and $h_3 = h_6 = 0$, so

$$\begin{aligned}\det\left(\nabla \underset{\mathtt{H}}{\boldsymbol{u}'}(\boldsymbol{u})\right) &= \det\left(\frac{I_2\left(\widetilde{\mathbf{h}}_3^\top\ \boldsymbol{u}+1\right) - \boldsymbol{u}\,\widetilde{\mathbf{h}}_3^\top}{(\widetilde{\mathbf{h}}_3^\top\ \boldsymbol{u}+1)^2}\right) \\ &= \det\left((h_7x+h_8y+1)^{-2}\begin{pmatrix}h_8y+1 & -h_8x \\ -h_7y & h_7x+1\end{pmatrix}\right) \\ &= (h_7x+h_8y+1)^{-3}. \end{aligned} \tag{4.8}$$

Setting $\det(\mathtt{A}) = \alpha^3$ and substituting equation 4.8 into equation 4.7 gives

$$s(\mathtt{H},\mathbf{u}) = \alpha^3\,(h_7x+h_8y+1)^{-3}, \tag{4.9}$$

and, after re-arranging to isolate the unknown quantities, the following linear constraint is obtained

$$\begin{pmatrix}x & y & -s(\mathtt{H},\mathbf{u})^{-\frac{1}{3}}\end{pmatrix}\begin{pmatrix}h_7 & h_8 & \alpha\end{pmatrix}^\top = -1. \tag{4.10}$$

4.2.3 Linear estimator for $\mathtt{H}_\infty$ based on local scale change

Members of LAF clusters are likely detections of the same repeating element in the scene pattern. Thus any LAF of the same cluster has the same scale in the scene plane, and is a measurement of the relative scale change $s(\mathtt{H},\mathbf{u})$ of the transformation $\mathtt{H}$ that maps the scene pattern to the imaged pattern. From equation 4.10, it is clear that three point locations and their corresponding scale changes $s(\mathtt{H},\mathbf{u}^{(i)})$ are required to estimate homography $\widetilde{\mathtt{H}}$. Recall from Section 2.4 that LAF $L^{(j)}$ is an 3-tuple of affine covariant points of the form

$$L^{(j)} = \left(\mathbf{u}_1^{(j)},\mathbf{u}_2^{(j)},\mathbf{u}_3^{(j)}\right) \quad \mathbf{u}_i^{(j)} = \begin{pmatrix}x_i^{(j)} & y_i^{(j)} & 1\end{pmatrix}^\top,$$

and the scale of a LAF is defined to be

$$\text{scale}(L^{(j)}) := \det\left(\mathtt{N}^{(j)^{-1}}\right) \quad \text{where } L_{\llcorner} = \mathtt{N}^{(j)}L^{(j)}.$$

Then given LAF $L^{(j)}$, the scale change of the region around the LAF $L^{(j)}$ with origin at $\mathbf{u}_2^{(j)}$ can be estimated as

$$s_L(\mathtt{H},\mathbf{u}_2^{(j)}) \approx \text{scale}(L^{(j)}). \tag{4.11}$$

To estimate $\widetilde{\mathtt{H}}$ given m LAFs, construct the design matrix $Z \in R^{m\times 3}$ as

$$Z = \begin{pmatrix}\vdots & \vdots & \vdots \\ x_2^{(j)} & y_2^{(j)} & -(\text{scale}(L^{(j)}))^{-1/3} \\ \vdots & \vdots & \vdots\end{pmatrix}, \tag{4.12}$$

where $x_2^{(j)}$,$y_2^{(j)}$ are the Euclidean coordinates of $L^{(j)}$'s origin, and solve

$$\begin{pmatrix}h_7 & h_8 & \alpha\end{pmatrix}^\top = -Z^\dagger \mathbf{1}^{m\times 1}. \tag{4.13}$$

Recovery of an $\mathtt{H}$ compatible with the constraints concatenated in Z is possible by choosing any affinity $\mathtt{A}$ such that $\det(\mathtt{A}) = \alpha^3$, and then setting $\mathtt{H} = \mathtt{A}\,\widetilde{\mathtt{H}}$. Since $\mathtt{H}$ transforms the scene plane so that matching LAFs become equiareal, $\mathtt{H}$ is a rectifying homography of the scene plane. The vanishing line l of the imaged scene plane is the pre-image of the line at infinity in the rectified frame $\begin{pmatrix}0 & 0 & 1\end{pmatrix}^\top$ [3]. Thus to recover the coordinates of the vanishing line of the scene plane, we apply the inverse of the rectifying line homography,

$$l = \mathtt{H}^\top \begin{pmatrix}0 & 0 & 1\end{pmatrix}^\top = \widetilde{\mathtt{H}}^\top \mathtt{A}^\top \begin{pmatrix}0 & 0 & 1\end{pmatrix}^\top = \widetilde{\mathtt{H}}^\top \begin{pmatrix}0 & 0 & 1\end{pmatrix}^\top = \begin{pmatrix}h_7 & h_8 & 1\end{pmatrix}^\top.$$

Now that the coordinates of the vanishing line $l = \begin{pmatrix}h_7 & h_8 & 1\end{pmatrix}^\top$ are known, any rectifying homography can be constructed with some affinity $\mathtt{A}$ as [11],

$$\mathtt{H}_\infty = \mathtt{A} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ h_7 & h_8 & 1 \end{pmatrix}.$$

4.2.4 Estimating $\mathtt{H}_\infty$ from many LAF clusters

In images with repetitive patterns, there are typically several detected LAF clusters. Additional LAFs introduce more constraints on the estimate of the vanishing line l, so their inclusion in the design matrix is beneficial. Since the LAFs are detected in the warped plane, the relative scale between sets of repeated elements in the scene plane cannot be accurately determined, so each added cluster requires the addition of a unique affine scale parameter α_k, and no constraint on the scale ratio between LAF clusters will be enforced. For simplicity, assume there are two detected LAF clusters: C_1, C_2. Then a linear system of equations is constructed and solved in a way analogous to equation 4.12,

$$\begin{pmatrix} \vdots & \vdots & \vdots & \vdots \\ x_2^{(C_1(i))} & y_2^{(C_1(i))} & -\left(\text{scale}(L^{C_1(i)}\right)^{-1/3} & \vdots \\ x_2^{(C_1(i+1))} & y_2^{(C_1(i+1))} & -\left(\text{scale}(L^{C_1(i+1)}\right)^{-1/3} & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ x_2^{(C_2(j))} & y_2^{(C_2(j))} & \vdots & -\left(\text{scale}(L^{C_2(j)}\right)^{-1/3} \\ x_2^{(C_2(j+1))} & y_2^{(C_2(j+1))} & \vdots & -\left(\text{scale}(L^{C_2(j+1)}\right)^{-1/3} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} h_7 \\ h_8 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}. \tag{4.14}$$

In summary, if there are m clusters and n LAFs, then there will be $m + 2$ parameters and n constraints. However, the expectation is that with any LAF cluster, $|C_k| \gg 2$.

The approximation for scale change $s(\mathtt{H}, \mathbf{u})$ was the result of the linearization given in equation 4.3. Thus the estimate of the inhomogeneous parameters of the vanishing line h_7, h_8 will have an error proportional to accuracy of $s(\mathtt{H}, \mathbf{u})$. In general, larger LAFs will result in higher error since the linearization is good only local to the LAF origin. The error can be reduced with an iterative approach that incrementally removes the perspective distortion of the LAFs and concatenates successive warps of partially-rectified LAFs to the estimate of $\mathtt{H}_\infty$ (see Algorithm 2). Looping stops when the estimated affine scale change is idempotent).

4.3 Motif and Repetition Estimation

Each LAF cluster represents inter-repetition tentative correspondences between pattern elements. It is expected that local neighborhoods surrounding clustered LAFs will contain large

Algorithm 2 Iterative linear estimate for $\mathtt{H}_\infty$

procedure ESTIMATE_$\mathtt{H}_\infty$($\text{LAFs}^{(0)}, \mathcal{C}$) ▷ Input is clustered LAFs
  $k \leftarrow 0$
  $\hat{\mathtt{H}}_\infty^{(0)} \leftarrow I_3$
  $\alpha_{j=1:m}^{(0)} \leftarrow 1$
  repeat
    $k \leftarrow k + 1$
    estimate $l^{(k)}$ and $\{\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_m^{(k)}\}$ from $\text{LAFs}^{(k-1)}, \mathcal{C}$ ▷ as in equation 4.14
    Construct $\hat{\mathtt{H}}^{(k)}$ from $l^{(k)}$
    $\text{LAFs}^{(k)} \leftarrow \hat{\mathtt{H}}^{(k-1)}\text{LAFs}^{(k-1)}$ ▷ rectify $\text{LAFs}^{(k-1)}$ with $\hat{\mathtt{H}}^{(k-1)}$
    $\hat{\mathtt{H}}_\infty^{(k)} \leftarrow \hat{\mathtt{H}}^{(k-1)}\hat{\mathtt{H}}_\infty^{(k-1)}$
  until $\max_i \left\{ \frac{\alpha_i^{(k)} - \alpha_i^{(k-1)}}{\alpha_i^{(k)}} \right\} < \epsilon$ ▷ Terminate when affine scale change is idempotent
  return $\hat{\mathtt{H}}_\infty^{(k)}$
end procedure

subsets of the pattern. LAFs of maximal clusters are used to define sets of overlapping neighborhoods in the pattern. Each LAF's neighborhoods is transformed to the canonical frame, where surrounding LAFs are aggregated. In the canonical frame, overlapping LAFs are identified as repeated elements and added to the motif. This process is repeated for a fixed number of the top LAF clusters. Mismatches are detected using spatial verification in the canonical frame. LAFs that do not overlap any other LAFs in the cluster are considered to be mismatched. The repetition with the most detected LAFs is chosen as the reference, and transforms $\mathtt{T}_i$ to all other repetitions are estimated. Subsequently, an initial guess of $\mathcal{M}$ is obtained by transforming each repetition $\mathcal{R}$ by $\mathtt{T}_i^{-1}$ to the reference and averaging corresponding LAFs from all repetitions.

4.3.1 Neighborhood definition

As aforementioned, LAF clusters represent a good starting point to begin construction of the pattern motif. A subset of maximal clusters $\mathcal{C}' \subset \mathcal{C}$ is selected to guide motif reconstruction. For each cluster $C_i \in \mathcal{C}'$, the origin of each member LAF $L^{(C_i)}$ is used to define the center of a search neighborhood. The radius ϵ of the neighborhood is the scaled estimated distance between LAF repeats in the rectified scene plane. The median is used to account for outliers,

$$\epsilon = \lambda \,\text{median}\left[\bigcup_k \left\{ \min_{j \neq k} d(\mathbf{u}_2^{C_i(j)} - \mathbf{u}_2^{C_i(k)}) \right\}\right] \quad \lambda > 1,$$

and the ϵ-neighborhood of LAF $L^{(C_i(j))}$ centered at LAF origin $\mathbf{u}_2^{C_i(j)}$ is denoted $N_\epsilon\left(L^{(C_i(j))}\right)$.

4.3.2 The canonical frame for LAF aggregation

A LAF cluster C_i defines the set of neighborhoods

$$\mathcal{N}_{C_i} = \bigcup_k \{N_\epsilon(L^{(C_i(k))})\},$$

which likely contain subsets of different pattern repetitions. Direct comparison between pattern repetitions is not possible because there remains an inter-repetition Euclidean ambiguity. Recall from Section 2.4 that for each LAF L, there is affinity $\mathtt{N}$ that transforms the LAF to the canonical frame. This normalization is used to canonicalize each neighborhood, $N_\epsilon(L^{(C_i(k))})$. The result is that the repeating elements of every neighborhood will have the same coordinates.

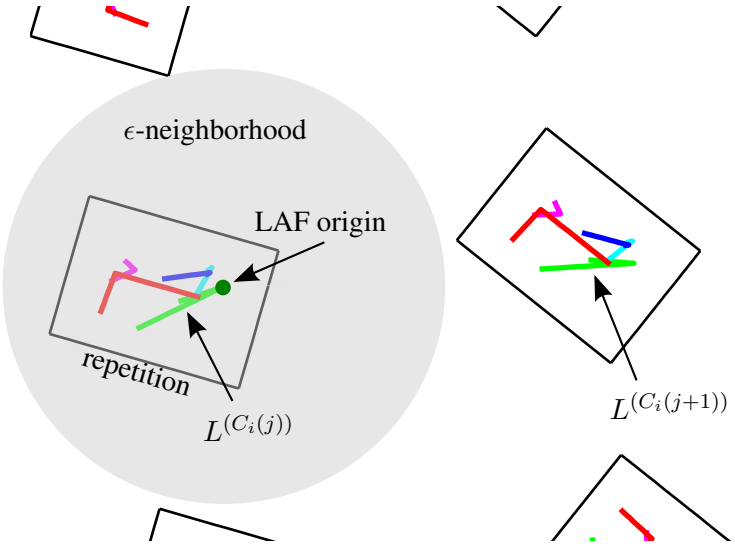


Figure 4.4 The ϵ-neighborhood as defined by the origin of LAF $L^{(C_i(j))}$. Pattern elements that are in the ϵ-neighborhood are tentatively considered to be part of one repetition. Tentative repetition membership is subsequently spatially verified.

4.3.3 Motif Construction

After neighborhood canonicalization, LAFs from different repetitions reside in the same space. Modulo detection noise, the expectation is that detected LAFs of the same repeating pattern element will overlap. The aggregated LAFs are agglomeratively clustered in the LAF space and a motif LAF is created for each spatial cluster.

Investigating detections across large data sets revealed that there are often multiple LAF detections in small neighborhoods of the repetitive pattern. Spurious detections can be quite close to the valid detections, *e.g.*, two of three LAF points might be the same scene element, but the third point differs between the two because of a detection error. For this reason, during clustering, the infinity norm is used $\|\cdot\|_\infty$, and a fairly tight threshold is chosen so that any deviation in any of the three LAF points from the rest of the cluster will result in its rejection.

For clustering, the inhomogeneous coordinates of the three LAF points are concatenated to create a 6-D representation for each LAF $L^{(i)}$,

$$\mathbf{l}^{(i)} = \begin{pmatrix} x_1^{(i)} & y_1^{(i)} & x_2^{(i)} & y_2^{(i)} & x_3^{(i)} & y_3^{(i)} \end{pmatrix}^\top.$$

The aggregated LAFs represented as 6-D vectors are complete-link agglomeratively clustered as described above. The criterion for clustering two LAFs $L^{(i)}$, $L^{(j)}$ is

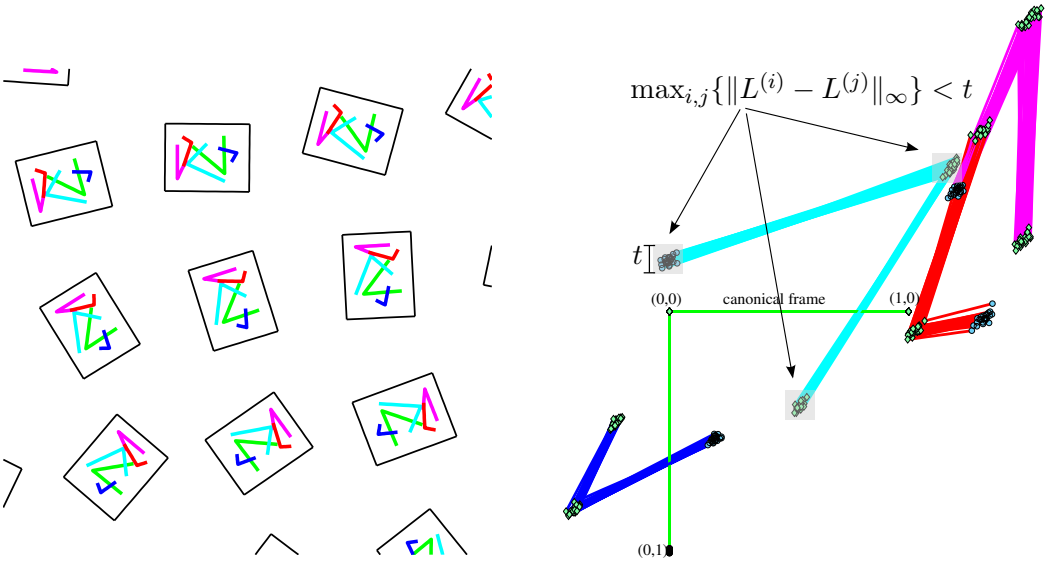
$$\{\|\mathbf{l}^{(i)} - \mathbf{l}^{(j)}\|_\infty\} < t \qquad t \text{ is a threshold in normalized pixels.}$$

Let $\{\mathcal{T}^{(k)}\}_{k=1}^m$ be the partition of the LAFs gotten by agglomerative clustering. Each $\mathcal{T}^{(k)}$ represents one repeating pattern element and $\cup_k \mathcal{T}^{(k)}$ are all LAFs in the pattern. The motif LAF corresponding to all all repetitions in $\mathcal{T}^{(k)}$ is estimated as

$$M^{(k)} = \frac{1}{|\mathcal{T}^{(k)}|} \sum_{L \in \mathcal{T}^{(k)}} L.$$

The motif is then just a tuple of motif LAFs,

$$\mathcal{M} = \left(M^{(1)}, M^{(2)}, \ldots, M^{(n)} \right).$$



(a) Affine rectified pattern. **(b)** Canonical frame for clustering LAFs.

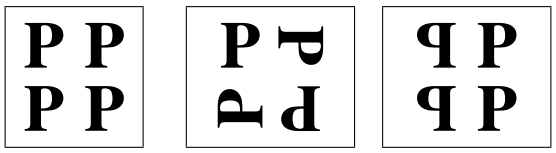
Figure 4.5 The green LAFs define the ϵ-neighborhood and the set of transforms to the canonical frame. For each repetition, LAFs in its ϵ-neighborhood are transformed to the canonical frame where they are spatially clustered. Clusters are subsequently used to estimate the motif. Note that feature detection noise, in this case $\sigma = 1.25$ pixels, prevents the LAFs from coinciding exactly.

With the partitonioning of repeated LAFs into their respective repititions, the set of inter-reptition transformations $\{\mathtt{T}_i\}_{i=1}^m$ can be estimated. Note that $\{\mathtt{T}_i\}_{i=1}^m$ are estimated from the affine rectified pattern. The type of the transform $\mathtt{T}_i$ determines the configuration of the pattern; *i.e.* whether repeitions of the motif are translated, rigidly transformed, reflected, or transformed by a general affinity. The next section will demonstrate that rigid transforms and reflections introduce constraints that can be used to further reduce the geometric ambiguity between the imaged pattern and scene pattern.

4.4 Reducing geometric ambiguity between the imaged pattern and scene pattern

In Algorithm 2, scene plane rectification is estimated from the ratio of corresponding LAF scales. The ratio is invariant to an affine transformation, hence the rectification is ambiguous up to an affine transformation. However, many real-world patterns have repeated scene elements that are not only equiareal, but also have corresponding lengths that are equal. In this section, it is shown that the length constraint reduces affine ambiguity down to a similarity (unknown scale and rotation) for certain configuratons of the repeated element. Specifically, if any repetition is rotated (Figure 4.6(b)), the ambiguity of the pattern rectification is reduced to a similarity transformation; if the pattern is reflected (Figure 4.6(c)), the ambiguity is reduced to a similarity plus an unknown scaling along the axis of symmetry; if the repeated pattern is only translated on the planar surface (Figure 4.6(a)), the ambiguity remains affine.

The following paragraphs describe the process of upgrading an affine-rectified pattern obtained from Algorithm 2 for different constructions of the repetitive pattern.



(a) (b) (c)

Figure 4.6 Examples of different configurations of a repetitive pattern: (a) Pure translation, (b) rigid transform, (c) axial symmetry (with translation).

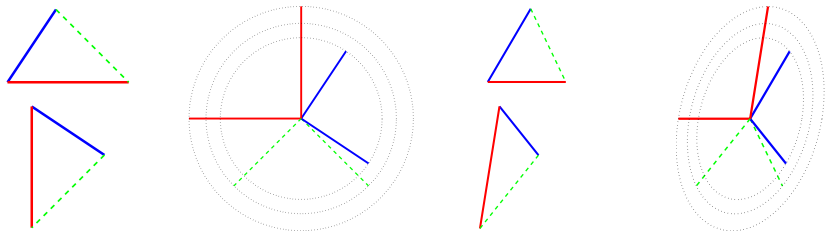


Figure 4.7 (a) repeated element is rotated and translated, (b) corresponding (color coded) vectors form circles, (c) repeated element transformed by an affine transformation, (d) corresponding vectors form ellipses that differ only by a diameter.

Pure translation. An affine transformation of a translated pattern can be also obtained as translated affinely transformed pattern

$$\begin{pmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & t'_x \\ 0 & 1 & t'_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.15}$$

$$\text{, where } \begin{array}{rcl} t'_x & = & a_1 t_x + a_2 t_y, \\ t'_y & = & a_4 t_x + a_5 t_y. \end{array}$$

Therefore, the length of the corresponding line segments remains equal under any affine transformation. Thus, a pattern repeated on a planar surface by pure translation can be only recovered up to an affine ambiguity, even if the length constraints are applied.

Rotation. Since the translation has no affect on the length of the line segment between points A and B, we will study the lengths of (free) vectors obtained from the line segment AB by translating A to the origin. Denote the vectors by $\mathbf{x} = (x, y)^\top = B - A$. The vectors will be indexed by two indices i and j. The index i is over the repetitions of the repeated element and j is over line segments within the repeated element. Vectors in the scene plane will be denoted $\hat{\mathbf{x}}_{ij}$. We will assume that all corresponding line segments (fixed index i) are of the same lengths

$$\hat{\mathbf{x}}_{ij}^\top \hat{\mathbf{x}}_{ij} = \hat{r}_i^2, \tag{4.16}$$

where r_i is the length of the line segments. The situation is depicted in Figure 4.7(b): corresponding vectors are color coded and Equation (4.16) is depicted as circles with diameter r_i. Let $\bar{\mathtt{A}}$ be the 2×2 upper left sub-matrix of an affine transformation $\mathtt{A}$. Transforming the plane by the affine transformation $\mathtt{A}$ transforms the vectors as $\mathbf{x}_{ij} = \bar{\mathtt{A}}\hat{\mathbf{x}}_{ij}$, and the length constraint (4.16) to

$$\mathbf{x}_{ij}^\top \Sigma^{-1} \mathbf{x}_{ij} = \hat{r}_i^2, \quad \text{where} \quad \Sigma = \bar{\mathtt{A}}\bar{\mathtt{A}}^\top. \tag{4.17}$$

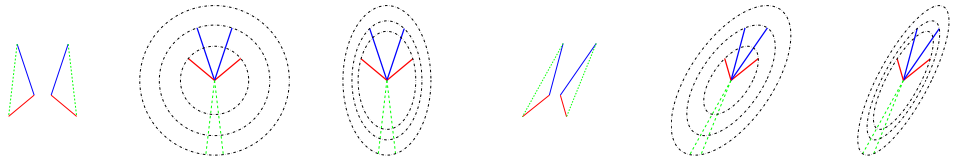


Figure 4.8 (a) pattern element repeated by reflection and translation, (b) and (c) corresponding (color coded) vectors form concentric ellipses with ambiguity in the scale along the axis of symmetry, (d) repeated element transformed by an affinity, (d) and (e) corresponding vectors form concentric ellipses with one degree of freedom.

In equation 4.17, Σ^{-1} represents an ellipse (visualized in Fig. 4.7c), where

$$\Sigma^{-1} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}. \tag{4.18}$$

Equation (4.17) can be rewritten as

$$(x_{ij}^2 \quad 2x_{ij}y_{ij} \quad y_{ij}^2 \quad -1)(a \quad b \quad c \quad r_i^2)^\top = 0. \tag{4.19}$$

This gives arise to a system of homogeneous linear equations. There are three unknowns for Σ^{-1}, and each set of matching line segments adds one unknown r_i. Each participating line segment in general position (rotation) adds one constraint. For two pairs of line segments, there are $3 + 2 = 5$ unknowns and four linear equations, yielding a one-dimensional linear space of solutions.

The affine transformation can be derived from the solution of the system of linear equations (4.19) up to a scale factor and a rotation. The unknown scale comes from the homogeneous nature of the system–both Σ^{-1} and r_i^2s can be multiplied by a positive scalar. The rotational ambiguity comes from the ambiguity of decomposition $\Sigma = \bar{\mathtt{A}}\bar{\mathtt{A}}^\top$.

A rotation by 180 degrees (or a integer multiple) creates a special case: if the pattern is only rotated by integer multiplications of 180 degrees, then the matching vectors lie on parallel lines. Since affine transformations affect the lengths of vectors on parallel lines equally, the situation is similar to the pure translation case with full affine ambiguity.

Symmetry. This paragraph examines the conifguration of a repeated element that is reflectied about a line of symmetry. This configuration frequently occurs on man-made objects, especially on building facades [34]. We will assume, without loss of generality, that the line of symmetry is a vertical line on the scene plane, see Figure 4.8(a). Expressing this in terms of matching vectors used in the previous paragraph, each set of matching vectors has only two distinct elements $\hat{\mathbf{x}}_\mathbf{1} = (\hat{x}_i, \hat{y}_i)^\top$ and from $\hat{\mathbf{x}}_2 = (-\hat{x}_i, \hat{y}_i)^\top$. Let $\bar{\mathtt{A}} = [\mathbf{a_1}\mathbf{a_2}]$. The observed vectors distorted by the affine transformation are

$$\mathbf{x}_{i1} = \bar{\mathtt{A}}\hat{\mathbf{x}}_\mathbf{1} = \hat{x}_i\mathbf{a_1} + \hat{y}_i\mathbf{a_2}, \quad \text{and} \tag{4.20}$$

$$\mathbf{x}_{i2} = \bar{\mathtt{A}}\hat{\mathbf{x}}_\mathbf{2} = -\hat{x}_i\mathbf{a_1} + \hat{y}_i\mathbf{a_2}. \tag{4.21}$$

Rewriting the length constraint (4.19), we get the two following equations,

$$(\hat{x}_i\mathbf{a_1} + \hat{y}_i\mathbf{a_2})^\top\Sigma^{-1}(\hat{x}_i\mathbf{a_1} + \hat{y}_i\mathbf{a_2}) = r_i^2, \tag{4.22}$$

$$(-\hat{x}_i\mathbf{a_1} + \hat{y}_i\mathbf{a_2})^\top\Sigma^{-1}(-\hat{x}_i\mathbf{a_1} + \hat{y}_i\mathbf{a_2}) = r_i^2. \tag{4.23}$$

Equations (4.22) and (4.23) arise from different sets of matching vectors (different index i) and are not linearly independent. Subtracting (4.22) from (4.23) we have

$$\mathbf{a_1}^\top\Sigma^{-1}\mathbf{a_2} + \mathbf{a_2}^\top\Sigma^{-1}\mathbf{a_1} = 0. \tag{4.24}$$

Thus, from the first set of matching vectors, we have $3+1$ (Σ^{-1} and r_1) unknowns and two linearly independent equations. Each additional set of matching vectors adds only one constraint of the form (4.22) and one unknown r_i. Therefore, the Σ^{-1} cannot be uniquely identified leading to a anisotropic scale ambiguity in the direction of the axis of symmetry. The ambiguity is depicted in Figures 4.8(b) and 4.8(c).

4.5 Summary

The goal of the single-view geometry developed in this section is to reduce the geometric ambiguity between imaged pattern and scene pattern, and to infer, through geometric constraints, the configuration of the pattern. Upgrades are applied to the imaged pattern so that it looks as close to possible to its scene counterpart. In the rectified plane, spatial relationships between LAF clusters are established from which the motif is estimated. The methods introduced require a minimal number of tentative correspondences between repeated LAF elements and they give accurate estimates across diverse camera geometries. By considering special configurations of the pattern, the ambiguity between scene and image can be reduced to a similarity.

5 Robust sparse 3-D reconstruction of repetitive patterns from single views

In Chapter 4, a series of linear estimates was introduced that, in aggregate, enable the sparse reconstruction of a imaged coplanar repetitive pattern. The estimates are robust across diverse camera geometries with the caveat that the camera doesn't deviate far from the pinhole model. Camera lenses can introduce non-linearities to the imaging geometry [11], which are not modeled in the single-view geometry already presented in Chapter 4. The presence of lens distortion can distort the pattern elements to the point that bootstrapping the geometry estimation fails, and such that the linear constraints on change of scale (see Section 4.2.1) are not valid. Further complicating geometry estimation is the problem of bad feature correspondence. In Chapter 3, it was shown how to establish tentative correspondences between the LAFs of a repeated element by matching SIFT descriptors. While SIFTs are discriminative, some of the LAF clusters will have mismatches. The linear estimators of single-view geometry introduced in Chapter 4 have a leverage of one point [13], so it is critical to exclude mismatches from geometry estimation to avoid a bad solution. The geometry pipeline developed to this point consists of a series of methods that linearly estimate geometry by minimizing algebraic errors. While the methods are effective, algebraic error has no physical meaning, and reprojection error is a much preferred objective for geometric and statistical reasons [11]. This chapter will introduce a unified and robust estimation framework to handle the multifarious issues enumerated above. The goal is to broaden the class of images in which pattern reconstruction is possible, and to achieve a statistically optimal estimate of the reconstructed pattern.

5.1 Overview

The challenges listed in the exposition motivate the design of the robust estimation framework. Two approaches are taken: (i) *random sampling*, which is used to test tentative correspondences and; (ii) *nonlinear estimation*, a "gold standard" algorithm which refines the linear estimate by estimating a full model with nonlinear constraint that minimizes geometric error. The methods from the Chapter 4 are used to give an initial guess to the refinement stage. The key distinction is that estimates are made from minimal subsets of data to globally probe the parameter space. Estimates from samples are used to verify tentative correspondences. The subset of good data is kept and an initial guess is made at the structure of the pattern (*i.e.* linear motif estimation and inter-repetition transforms). The initial guess is re-parameterized with *nonlinear* constraints and added to the parameterization is the lens distortion coefficient. The resulting system of nonlinear equations is solved by minimizing the geometric reprojection error of the reconstructed pattern to the detected LAFs.

5.2 Verifying LAF correspondences with RANSAC

Recall from Chapter 3 that SIFT clustering was used to establish correspondences between pattern repeats. Because of various photometric effects and geometric distortions, mismatches can be included into the sets of correspondences. Recall from Chapter 4 that the removal of geometric ambiguity between the imaged pattern and scene pattern begins with affine rectification

of the imaged pattern. In Section 4.2.3 a method for estimation of $\mathtt{H}_\infty$ from 3 LAF correspondences is presented. The method is not tolerant to bad correspondences, though. A principled method for sampling the correspondences for good matches is needed.

Because of its exceptional tolerance to highly corrupted data, The Random Sampling Consensus RANSAC [8] algorithm is commonly used to discard mismatched correspondences. RANSAC is a hypothesize and test framework: the minimal sample size is repeatedly drawn (as defined by the model) until either a sufficient number of good data are verified or a sufficient number of samples have been drawn. Particular to the problem at hand, a way to define good data is needed. After affine rectification, repeated elements have the same scale in the image, so the pair-wise scale ratios of intra-cluster rectified LAFs can be thresholded to determine the good correspondences. The number of good correspondences defines the RANSAC score of the hypothesized rectifying homography. Let $\{L_{\mathtt{H}}^{C_k(i)}\}_{i=1}^{m}$ be a set of rectified LAFs. Then the number of matched LAFs is in the cluster is computed as

$$\text{score} := \sum_k \max_j \left\{ \left| \left\{ i \colon t_1 < \frac{\text{scale}\left(L_{\mathtt{H}}^{C_k(i)}\right)}{\text{scale}\left(L_{\mathtt{H}}^{C_k(j)}\right)} < t_2 \right\} \right| \right\} \quad \text{where } t_1 < 1 < t_2 \tag{5.1}$$

The hypothesis with the best score as defined in Equation 5.1 is kept and all LAFs that were counted in the score are considered verified. Algorithm 2 (with the modification for multiple LAF clusters as described in Section 4.2.4) is used for the hypothesis generation. LAF clusters are uniformly sampled based on their cardinality and a subset of 3 LAFs are chosen from each sampled cluster. The LAFs are used to generate a rectifying hypothesis, and its quality is assessed by the scoring function given in Equation 5.1. After estimation of the initial rectifying homography and verification of LAFs the geometry pipeline continues as developed in Chapter 4.

5.3 Lens Distortion

The assumption through the development in Chapter 4 is that the pinhole model [11] is an accurate model for image capture. Tacit in this assumption is that scene lines project to image lines. For many classes of lenses, this approximate model is not valid because of the effects of *lens distortion*. The most commonly used lens distortion model is the radial distortion model, named so because the magnitude of distortion is a function of the distance to the center of distortion [11]. In particular, we use the one-parameter division model [9] for lens distortion. The distortion center is assumed to be known; the lens distortion parameter is denoted by λ; and the distortion function for particular λ is denoted $\mathcal{L}_\lambda$ and undistortion denoted $\mathcal{L}_\lambda^{-1}$.

If the lens distortion is significant, LAFs can undergo significant distortion, especially at the periphery of the image. But this presents a "chicken and egg" problem. To estimate the lens distortion λ, LAF correspondences are needed (the presence of straight lines is not assumed); conversely, to achieve reliable LAF correspondence, the lens distortion is needed. To account for significant lens distortions, we coarsely sample an interval of lens distortions typical for commonly used lenses. The robust estimation framework is run for each sample, and the estimate with the smallest root mean square re-projection error is kept.

5.4 Nonlinear estimation

In this section, we formulate the process of reconstructing a coplanar repetitive pattern captured by a perspective camera *with* lens distortion. This algorithm reconstructs patterns that

can be generated by applying to their motif a series of transformations $\{\mathtt{T}_i\}_{i=1}^m$ from the following geometric strata: pure translation; Euclidean (translation, rotation and reflection); and, less commonly, affine. In principle, the transformations $\{\mathtt{T}_i\}_{i=1}^m$ could be projective, but it is impossible to decouple the projective effects of the camera and projectivities, which makes estimation of the transformations $\{\mathtt{T}_i\}_{i=1}^m$ impossible. Thus we consider only patterns that can be modeled with non-projective transformations $\{\mathtt{T}_i\}_{i=1}^m$, which still includes virtually all man-made repetitive patterns.

Estimation of the motif $\mathcal{M}$, inter-repetition transformations $\{\mathtt{T}_i\}_{i=1}^m$ and partial camera calibration $\mathtt{H}_\infty$ is accomplished by the linear pipeline developed in Chapter 4. Rotated repetitions, if present, are detected in the affine rectified pattern and the length constraint discussed in Section 4.4 is used to estimate affinity $\hat{\mathtt{A}}$, which, when left-multiplied to the estimated rectifying homography $\hat{\mathtt{H}}_\infty$, lifts the scene plane from the projective to the Euclidean stratum. The 3-tuple $\left(\hat{\mathtt{H}}_\infty, \hat{\mathcal{M}}, \{\hat{\mathtt{T}}_i\}_{i=1}^m\right)$ is the initial guess provided to the non-linear solver.

5.4.1 Parameterization

The parameterization of inter-repetition transformations is contingent on the pattern configuration detected by methods introduced in Section 4.4. Regardless of configuration, the repetition with the most feature detections is chosen to anchor the coordinate system. All inter-repetition transforms are estimated relative to the anchor repetition. Depending on the configuration detected, one of the following parameterizations of $\{\mathtt{T}_i\}_{i=1}^m$ is adopted:

translation $\{\mathtt{T}_i\}_{i=1}^m = \left\{\begin{pmatrix} dx^{(i)} & dy^{(i)} \end{pmatrix}\right\}_{i=1}^{m-1}$, translation parameters. Inter-repetition translations resemble their real-world counterparts up to an affinity.

rigid transform $\{\mathtt{T}_i\}_{i=1}^m = \left\{\begin{pmatrix} dx^{(i)} & dy^{(i)} & \theta^{(i)} \end{pmatrix}\right\}_{i=1}^{m-1}$, translation and rotation parameters. This parameterization is chosen if any rotation is detected among the repetitions. Inter-repetition rigid transforms resemble their scene counterparts up to a similarity.

reflection or affine $\{\mathtt{T}_i\}_{i=1}^m = \left\{\begin{pmatrix} a_{11}^{(i)} & a_{12}^{(i)} & a_{13}^{(i)} & a_{21}^{(i)} & a_{22}^{(i)} & a_{23}^{(i)} \end{pmatrix}\right\}_{i=1}^{m-1}$, affine matrix parameters. The current implementation does not take advantage of constraints introduced by reflections, but if a reflection is present, inter-repetition transforms resemble their scene-plane counterparts up to similarity with an anisotropic scale ambiguity.

Shum and Szeliski [29] introduce an 8 parameter model of the homography that gives better convexity in the error surface than direct parameterization for non-linear estimation,

$$\hat{\mathtt{H}}_\infty^* = \begin{pmatrix} 1+\Delta h_1 & \Delta h_2 & \Delta h_3 \\ \Delta h_4 & 1+\Delta h_5 & \Delta h_6 \\ \Delta h_7 & \Delta h_8 & 1 \end{pmatrix} \begin{pmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{pmatrix}. \tag{5.2}$$

Constants $\{h_i\}_{i=1}^9$ are the individual entries of $\hat{\mathtt{H}}_\infty$ obtained from Algorithm 2. Parameters $\{\Delta h_i\}_{i=1}^8$ are to be estimated. Recall from Section 5.3 that the one-parameter (λ) division model is used for modeling radial lens distortion. An initial guess at λ is obtained by uniformly sampling an interval of lens distortion values common to consumer lenses. Lens distortion is parameterized as a perturbation to this initial guess,

$$\hat{\lambda}^* = \lambda_0 + \Delta\lambda \quad \lambda_0 \text{ coarsely sampled} \tag{5.3}$$

The motif is parameterized as displacements to the positions of the member LAFs estimated in Section 4.3.3. This implies that given n motif member LAFs, $6n$ displacements are added to the parameterization (2 for each affine covariant LAF point).

5.4.2 Optimization

Non-linear estimation proceeds by minimizing the geometric re-projection error of the reconstructed pattern with its imaged counterpart. In the non-linear optimization, we seek optimal lens distortion $\hat{\lambda}^*$, rectifying homography $\hat{\mathtt{H}}^*_\infty$, motif $\hat{\mathcal{M}}^*$, and inter-repetition transforms $\hat{\mathtt{T}}^*_i$ that minimize the re-projection error,

$$\min_{\hat{\lambda},\hat{\mathtt{H}}_\infty,\hat{\mathtt{T}}_i,\hat{\mathbf{v}}_k^{(i)}} \sum_{C_i\in\mathcal{C}^*}\sum_{j\in C_i}\sum_{k=1}^{3} d(\mathcal{L}_{\hat{\lambda}}(\hat{\mathtt{H}}_\infty^{-1}\hat{\mathtt{T}}_{\mathtt{j}}(\hat{\mathbf{v}}_k^{(i)})), \mathbf{u}_k^{C_i(j)})), \tag{5.4}$$

where $\hat{\mathbf{v}}$ is the set of affine covariant points of the member LAFs of the estimated motif $\hat{\mathcal{M}}$, and $d(\mathbf{x},\mathbf{y})$ is the Euclidean distance between the images of $\mathbf{x}$ and $\mathbf{y}$.

5.5 Synthetic Tests

To gain a principled understanding of the usefulness of the framework, performance was analyzed with synthetic cameras and patterns. The virtual scene is constructed by generating a coplanar repetitive pattern in 3-D and placing a camera at a distance from the scene plane such that the pattern occupies most of the view. The motif, spacing and orientation of the repetitions are generated by random rigid transforms. The camera intrinsics, modulo radial distortion, are uniformly sampled from values that are typical for video and film imagery. The position of the camera varies uniformly on the upper $2/3$ of a hemisphere with the scene plane as the equator. The camera is oriented such that the principal ray intersects a point drawn from a Gaussian distribution on the scene plane.

Performance is measured against two parameters: feature detection noise σ, and radial lens distortion parameter λ. The quantities are independently, uniformly sampled over intervals of values commonly found in consumer cameras. Algorithm 3 outlines the test process. The noise

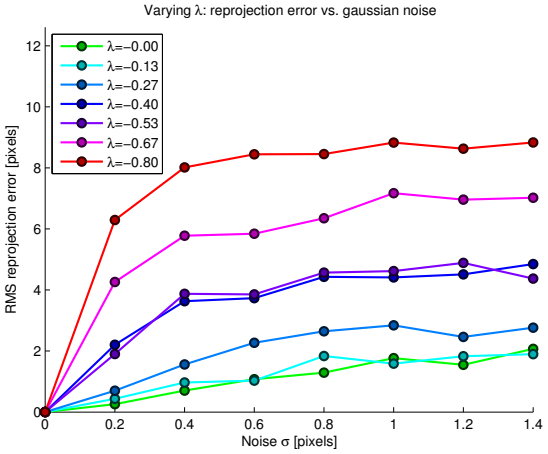
Algorithm 3 Performance evaluation over varying feature detection noise and lens distortion.

1. Repeat m times
   a) Randomly generate a 3-D pattern and camera. Project the pattern into the image plane to get the *perfect* image points.
   b) For each value of σ and λ, repeat n times.
      i. Radially distort the perfect points by λ.
      ii. Draw noise from a Gaussian distribution of standard deviation σ and add to projected points.
      iii. Sparsely reconstruct pattern from the distorted noisy points and estimate the radial distortion parameter $\hat{\lambda}$.
      iv. Reproject reconstructed 3-D points into the image plane and compute the root-mean square distance (RMS) between the reprojected reconstructed pattern and the corresponding perfect points.
2. From the mn runs of each (σ,λ) pair, root mean squared (RMS) distances and estimated lens distortions $\hat{\lambda}$ are plotted. For each λ the median, 20th and 80th percentile estimates of lens distortion $\hat{\lambda}$ is plotted against all generated noise values σ. See Figure 5.1.

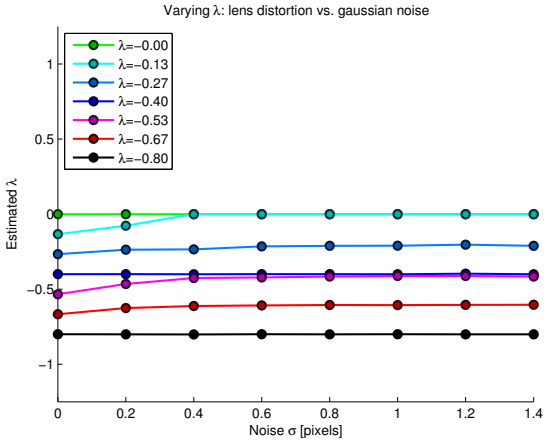
levels were sampled from $\sigma \in [0, 1.4]$ and the lens distortion sampled from $\lambda \in [-0.8, 0]$. most cameras common cameras have a σ of about 0.2 pixels after sub-pixel interest-point extraction [9], so the algorithm is tested against exceedingly high noise levels. Seven virtual scenes were randomly generated ($m = 7$ in Algorithm 3) and for each (σ, λ) pair, five repetitions ($n = 5$ in Algorithm 3) were made to measure over different noise generations and to account for the randomized nature of the estimation framework (in particular RANSAC). Test results are summarized in Figure 5.1.

5.5.1 Analysis

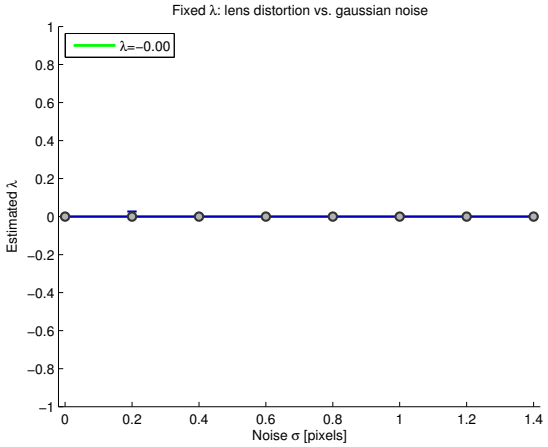
Examining Figure 5.1 allows a number of conclusions about the robust framework to be drawn. Firstly, the RMS error increases nearly linearly with increasing noise at smaller distortion levels. At typical noise levels and lens distortion levels, RMS reprojection errors are less than a pixel. At the higher distortion levels, where the displacement between undistorted and distorted points can be $\approx 10^2$ pixels, RMS makes an initial jump, but remains stable across high noise levels. With high noise levels and high distortions, there will be some coupling of the effects, so the behavior is expected. Secondly, there is a systematic bias in the estimate toward less extreme distortions (greater λ) as noise level increases. Curiously, this is an opposite finding to what Fitzgibbon found for his lens distortion estimation framework for the division model [9]. For near noiseless cases, the algorithm is very stable at estimating λ, suggesting a wide-basin of convergence. For most λ at all noise levels, including those representing much higher levels than encountered with feature detectors, lens distortion λ was correctly estimated. An interesting phenomenon occurs at $\lambda \in \{0.13, 0.67\}$: the estimates are consistently biased to lower distortions. This might be explained by the coarse sampling used to give a rough initial guess at the lens distortion. The initial guess is likely pushing the algorithm toward a local optimum. Overall the results are encouraging. The algorithm accurately reconstructs the pattern in the presence of significant lens distortions and at high noise levels.



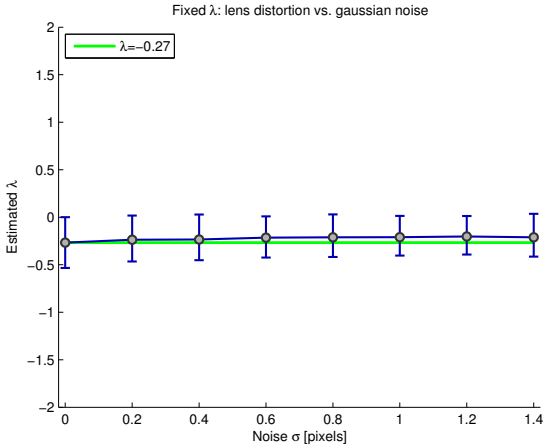
(a) median error for each (σ, λ) pair.



(b) median $\hat{\lambda}$ versus increasing noise.



(c) median $\hat{\lambda}$, fixed $\lambda = 0.00$.



(d) median $\hat{\lambda}$, fixed $\lambda = -0.27$.

Figure 5.1 Stability of the robust estimation framework as measured with 1960 virtual scenes generated as outlined in Algorithm 3: (a) by the median root mean square (rms) reprojection error; (b) by the median estimated λ; (c) for fixed $\lambda = 0.00$, the median and 20 percentile error bars are plotted; (d) as in (c), but with $\lambda = -0.27$.

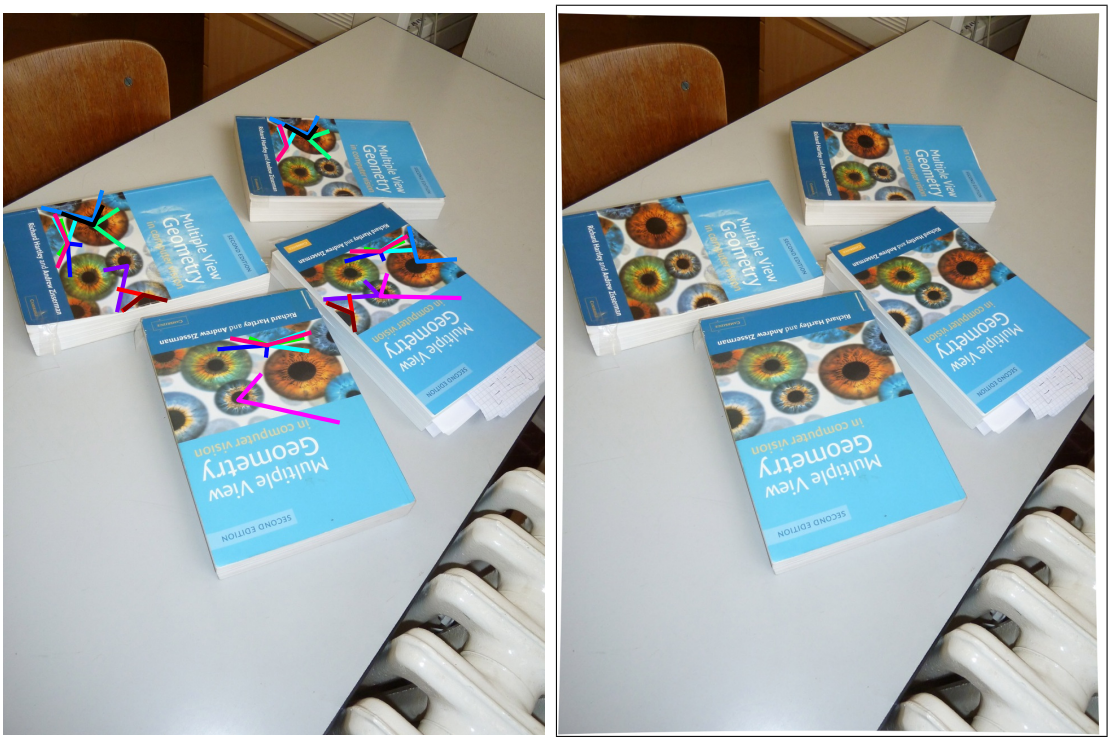
5.6 Results on real data

Results on three representative pattern configurations are presented: translations only, rotations only and rigid transforms. All images were acquired with cameras that exhibited some lens distortion. An additional result is presented on a building facade acquired by a lens with significant lens distortion. The same parameters were used for each result, demonstrating the automated nature of the algorithm and its invariance to small changes in thresholds. The reconstructed pattern is reprojected into the original image. Color correspondence indicates the same repeated element in the reconstruction (they should correspond one-to-one to repeated elements in the scene pattern).

(a) reprojected reconstructed pattern **(b)** undistorted image

(c) undistorted rectification

Figure 5.2 Translation configuration. The window arrangement gives the repetitive pattern. A rotation is not present in the repetitive pattern, so only affine rectification is possible. Radial distortion is estimated, which can be observed, in particular, on the straightened flagpole in the undistorted images. Root mean square reprojection error of the reprojected reconstruction is 0.604 pixels.



(a) reprojected reconstructed pattern

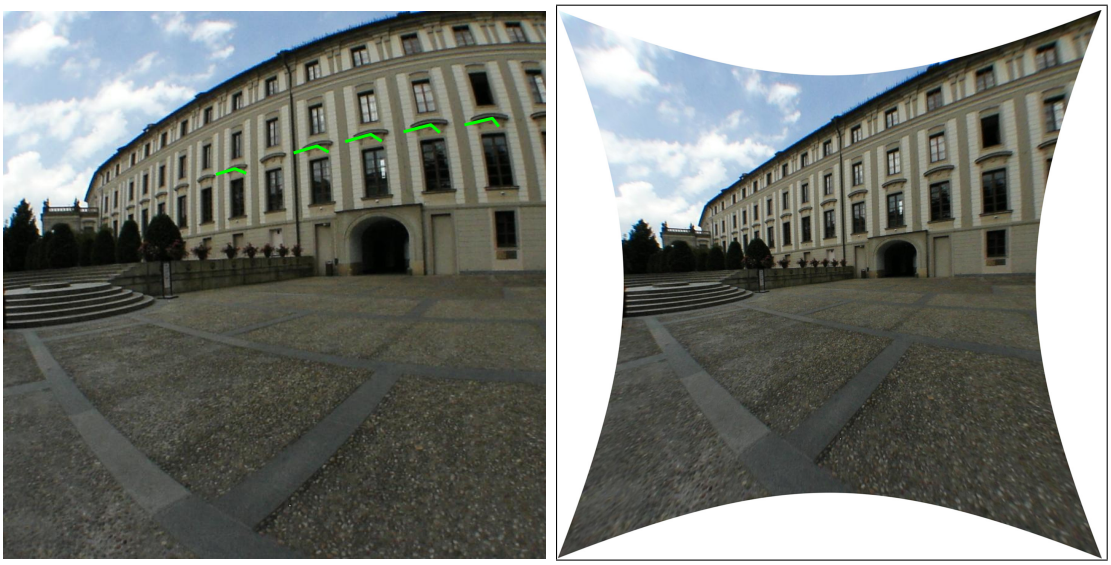
(b) undistorted image

(c) undistorted rectification

Figure 5.3 Arbitrary rigid transform configuration. The book covers give the repetitive pattern. A rotation is detected in the repetitive pattern, so geometric ambiguity between the imaged patter and scene pattern is reduced to a similarity. Root mean square reprojection error of the reprojected reconstruction is 1.65 pixels.

(a) reprojected reconstructed pattern **(b)** undistorted image

Figure 5.4 Image from *Planar Affine Rectification from Change of Scale* [3]. Chum et al. successfully rectify the pattern, but do not undistort the image. Notice in the undistorted image that straight scene lines are imaged straight. Additionally, the pattern is reconstructed within a similarity transform of the dart board in the scene. Shown is the reprojected pattern.



(a) reprojected reconstructed pattern **(b)** undistorted image

Figure 5.5 Severe lens distortion. Because of the severe lens distortion and oblique angle of the facade, only one repeated element of the window array was detected. Still, the robust estimation framework successfully reconstructed the pattern (as seen by its reprojection in the original image), and accurately estimated the lens distortion (straight scene lines project to straight imaged lines) despite the sparse data. This re-affirms the performance characterization of the synthetic tests. Root mean square reprojection error of the reprojected reconstruction is 2.75 pixels.

6 Conclusions

A novel, robust and statistically optimal framework has been presented to sparsely reconstruct imaged coplanar repetitive patterns. The framework is an end-to-end system: it begins by ingesting images and outputs the 3-D reconstructed pattern as a motif with inter-repetition transforms that, together, define the pattern configuration. To broaden the class of images to which the method is applicable, radial lens distortion is also estimated as a nuisance parameter, making the algorithm a candidate for automated camera calibration as well.

To realize the algorithm, a novel set of linear single-view geometric constraints was introduced. Linear estimators were derived from the constraints for the purpose of reducing the geometric ambiguity between the imaged plane and scene plane so that the configuration of the pattern could be inferred. These linear estimators require a minimal set of correspondences, and are effective with a diverse set of camera geometries. Robustness to errors in inter-pattern feature matching was obtained by wrapping the linear pipeline in a RANSAC loop. A statistically optimal estimator was designed to concurrently minimize geometric error and estimate non-linear radial lens distortion. Synthetic tests on an wide variety of virtual scenes affirmed the robustness and accuracy of the framework. Selected results on real image data were also presented that demonstrate the algorithm works on varied scene content and arbitrary pattern configuration.

The goal of automated detection and modeling of imaged coplanar patterns has been achieved. Future directions include the integration of this framework into an image retrieval system to determine the benefit that repetitive pattern modeling brings.

Bibliography

[1] M. Antone. Robust camera pose recovery using stochastic geometry, 2001. 8

[2] G. J. Burghouts, A. W. M. Smeulders, and J. M. Geusebroek. The distribution family of similarity distances. In *Advances in Neural Information Processing Systems*, volume 20, 2007. 15

[3] O. Chum and J. Matas. Planar affine rectification from change of scale. In *ACCV*, 2011. 6, 18, 19, 22, 37

[4] F. Cole, K. Sanik, D. DeCarlo, A. Finkelstein, T. Funkhouser, S. Rusinkiewicz, and M. Singh. How well do line drawings depict shape? *ACM Trans. Graph.*, 28(3):28:1–28:9, July 2009. 8

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886 –893 vol. 1, june 2005. 10

[6] P. Doubek, J. Matas, M. Perdoch, and O. Chum. Image matching and retrieval by repetitive patterns. In *ICPR*, 2010. 6

[7] R. Fabbri and B. Kimia. 3d curve sketch: Flexible curve-based stereo reconstruction and calibration. In *CVPR*, pages 1538–1545, 2010. 8

[8] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *CACM*, 24(6):381–395, June 1981. 18, 30

[9] A. W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *CVPR*, pages 125–132, 2001. 30, 33

[10] A. François, G. Medioni, and Waupotitsch R. Mirror symmetry ⇒ 2-view stereo geometry. *Image and Vision Computing*, 21:137–143, 2003. 6

[11] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 18, 22, 29, 30

[12] W. Hong, A. Yang, K. Huang, and Y. Ma. On symmetry and multiple-view geometry: Structure, pose, and calibration from a single image. *IJCV*, 60(3):241–265, December 2004. 6, 7

[13] P.J. Huber. *Robust Statistics*. 15, 29

[14] H. Jegou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*, pages 1169–1176, 2009. 4

[15] Y. Liu, R. Collins, and Tsin Y. A computational model for periodic pattern perception based on frieze and wallpaper groups. *PAMI*, 26:354–371, 2004. 6

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 4, 10, 13

[17] U. Luxburg. A tutorial on spectral clustering. *CoRR*, abs/0711.0189, 2007. 13

[18] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, pages 384–393, 2002. 8

[19] J. Matas, S. Obdržálek, and O. Chum. Local affine frames for wide-baseline stereo. In *ICPR*, volume 4, pages 363–366, 2002. 8, 9, 10

[20] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. 2004. 10

[21] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 1(60):63–86, 2004. 8, 9

[22] S. Obdržálek and J. Matas. Object recognition using local affine frames on maximally stable extremal regions. 8

[23] Š. Obdržálek and J. Matas. Object recognition using local affine frames on distinguished regions. pages 113–122, 2002. 9

[24] M. Park, R. Collins, and Liu Y. Deformed lattice discovery via efficient mean-shift belief propagation. *ECCV*, pages 474–485, 2008. 6

[25] M. Perdoch, J. Matas, and S. Obdržálek. Stable affine frames on isophotes. 2007. 8

[26] Arandjelovic R. and Zisserman A. Three things everyone should know to improve object retrieval. In *CVPR*, pages 2911–2918, 2012. 10, 11

[27] F. Schaffalitzky and A. Zisserman. Geometric grouping of repeated elements within images. In *BMVC*, pages 13–22. Springer-Verlag, 1998. 6, 7

[28] G. Schindler, P. Krishnamurthy, R. Lublinerman, Y. Liu, and F. Dellaert. Detecting and matching repeated patterns for automatic geo-tagging in urban environments. In *CVPR*, pages 1–7, 2008. 6

[29] H. Shum and R. Szeliski. Construction of panoramic image mosaics with global and local alignment. *International Journal of Computer Vision,36(2):101*, 2000. 31

[30] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, 2003. 4, 13

[31] Leung T. and Malik J. Detecting, localizing and grouping repeated scene elements from an image. In *ECCV*, pages 546–555. Springer-Verlag, 1996. 6

[32] T. Tuytelaars, A. Turina, and L. Van Gool. Noncombinatorial detection of regular repetitions under perspective skew. *PAMI*, 25(4):418–432, April 2003. 6, 7

[33] L. Van Gool, M. Proesmans, and A. Zisserman. Planar homologies as a basis for grouping and recognition. *Image and Vision Computing*, 16(1):21–26, 1998. 6

[34] C. Wu, J.-M. Frahm, and M. Pollefeys. Detecting large repetitive structures with salient boundaries. In *ECCV*, pages 142–155, Berlin, Heidelberg, 2010. Springer-Verlag. 6, 7, 27

[35] C. Wu, J. M. Frahm, and M. Pollefeys. Repetition-based dense single-view reconstruction. In *CVPR*, pages 3113–3120, 2011. 6, 7