Geometric Image Segmentation via Multiscale TILT Clustering

Chi Pang Lam
Allen Yang
Ehsan Elhamifar
S. Shankar Sastry

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2013-132
http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-132.html

July 16, 2013

Acknowledgement

This work was supported in part by DARPA FA8650-11-1-7153 and ARO 63092-MA-II.

Geometric Image Segmentation via Multiscale TILT Clustering*

Chi-Pang Lam, Allen Y. Yang, Ehsan Elhamifar, S. Shankar Sastry
Department of EECS, University of Calnifornia, Berkeley
{cplam,yang,ehsan,sastry}@eecs.berkeley.edu

Abstract

We present a novel algorithm to acquire and analyze rich 3D geometric features in single urban images. Traditional representation of 3D structures via local image features lack global geometric information to provide high-quality image correspondence and 3D models. The new approach utilizes the low-rank representation technique to seek a new class of invariant features based on minimizing the matrix rank of image textures, which are more holistic with respect to global geometric information, invariant to camera distortion, and robust to pixel corruption. Based on the transform-invariant low-rank texture (TILT) representation, we first propose an efficient algorithm to detect TILT features in urban images where man-made, symmetric patterns are abundant. Second, we introduce a multiscale, top-down representation of TILT clusters as TILT complexes, each of which represents a dominant planar structure (e.g., building facades) in 3D space. Extensive experiments are conducted on the Pankrac building database to demonstrate the efficacy of the algorithm. The source code of the algorithm will be available for peer evaluation.

1. Introduction

In computer vision, it has been well known that traditional image features such as corner points and edges do not contain sufficient 3D geometric information *alone*. As a result, inferring 3D geometry using these basic features on single or multiple images has been a difficult *inverse problem*, partly because the global geometric relationship between 3D shapes in space has been "destroyed" during the feature extraction stage. Furthermore, the basic image features extracted from local image pixels can be easily affected by many image nuisances such as illumination change, camera perspective projection, and occlusion. Therefore, it is desirable in many vision applications to instead extract image features that contain richer semantic or geometric information, whose representation as vectors or matrices is invariant to those image nuisances. In general, this category of robust image features are known as *invariant features*.

In the literature, many types of invariant features have been proposed. Arguably the most influential ones are the affine-invariant SIFT features and many of its variants [15, 18, 19, 1]. Since point and line features used in traditional *structure-from-motion* (SfM) approaches are not invariant to camera transformation and illumination, SIFT-type features expand the representation of image appearance to a small local window and consider the distribution of its pixel values and gradients. In urban-scene modeling, symmetric texture regions are also widely used [29, 14, 5]. Using the virtual views of symmetric patterns, their 3D orientation can be readily estimated from just a single image [12, 13]. Another type of geometric features used in 3D modeling are homogeneous color regions such as superpixels [22] whose orientation under perspective projection is consistent with that of some global planar structures in space [20, 25]. Finally, in object recognition and segmentation, various types of object part-based regions that contain rich semantic information have been proposed [33, 10, 27].

More recently, motivated by the emerging theory of Robust PCA [4], a new type of invariant feature has been proposed, called *transform-invariant low-rank texture* (TILT) [32]. The fundamental idea of TILT is that image texture that represents regular or repetitive 3D shapes in space is often low rank, when the texture region is represented as a matrix of its pixel values. However, under camera perspective distortion and potential pixel corruption, the matrix representation of the texture in the image space exhibits much higher rank compared to its *canonical representation*, i.e., the texture observed under orthographic projection and free of pixel corruption. Therefore, the rank of the texture region can be used as part of an objective function to rectify the underlying image distortion. This new approach suggests that we can obtain accurate geometric models of many urban objects, such as buildings, hallways, road signs, and human faces, without relying on extraction of any traditional local features (as shown in Figure 1). More importantly, the resulting TILT features can be shown to be robust to camera

*This work was supported in part by DARPA FA8650-11-1-7153 and ARO 63092-MA-II.

perspective distortion and can also compensate a moderate amount of pixel corruption, which are the main advantages of the method compared to other existing invariant features.

Figure 1: Examples of manually labeled image patterns that are extracted as TILT features. **Top:** Initialization of the feature locations as the red bounding boxes, and the final orientation of the feature as the green bounding boxes. The TILT features compensate the perspective distortion. **Bottom:** Canonical representation of the low-rank matrices.

Despite attractive attributes of TILT, it has not been widely adopted in many vision application where the use of invariant features would be preferred. We are aware of three applications in the existing literature: 3D reconstruction of building facades [21], symmetry detection [31], and camera calibration [30]. Compared to a typical natural image where the presence of low-rank texture may be only sporadic, the images used in the above applications mostly have overwhelming regular and/or repetitive patterns. However, the detection of TILT features in the previous work was achieved either by user input or by applying a fixed grid on the images.

1.1. Contributions

In this paper, we propose a novel algorithm to address two critical issues that have handicapped the use of TILT features in vision applications. First, we propose a simple yet effective algorithm to detect low-rank texture regions in natural images. It also effectively rejects texture/color regions that do not contain useful geometric information of the scene, e.g., the texture of bushes or sky.

Second, after extracting a set of TILT features from an image, we further propose a principled solution to partition the features into groups, each of which represents a unique 3D planar structure. More specifically, we build a 2D adjacency graph, where each node in the graph corresponds to a TILT feature. We connect two adjacent features by an edge whose associated weight is derived from their low-rank representations. As some of the nodes in the graph correspond to outlying features, in order to cluster the graph while rejecting the outlying nodes, we employ a recent robust clustering algorithm proposed in [7].

Different from most other image segmentation algorithms, the new segmentation algorithm is based on the robust canonical representation of image texture measured by its matrix rank. At the end, given a natural image as the input, the result of the algorithm provides a *geometric segmentation* of the image scene into regions with consistent 3D orientation and surface texture, as shown in Figure 2. The segmentation result can be readily employed by other higher-level algorithms in object recognition, image retrieval, and 3D reconstruction, to just name a few. Finally, to aid peer evaluation, the source code of our algorithm in MATLAB will be made available on our website.

2. Problem Formulation

In this section, we first review the basic TILT framework. Suppose $A \in \mathbb{R}^{m \times m}$ represents the image of a low-rank texture pattern, which can be distorted by a 3D transformation τ and sparse pixel corruption $E \in \mathbb{R}^{m \times m}$.[1] Therefore, under such transformation τ, the relationship between the distorted input image I and its ground-truth low-rank component A can be modeled as:

$$I \circ \tau = A + E. \tag{1}$$

In a sense, the appearance of a grayscale image patch I treated as a matrix can be decomposed as $I = (A, E, \tau)$, where τ is camera projection, E is a sparse pixel corruption matrix, and A is a low-rank texture pattern invariant to τ and E. We refer A as a *canonical representation* of I. In this paper, we restrict our attention to model planar texture patterns. Hence, τ is assumed to belong to the *homography group* $GL(3)$.

Motivated by the Robust PCA algorithm [4], (A, E, τ) can be recovered by solving the following optimization program:

$$\min_{A,E,\tau} \|A\|_* + \lambda \|E\|_1 \quad \text{subj. to} \quad I \circ \tau = A + E, \tag{2}$$

where $\|\cdot\|_*$ and $\|\cdot\|_1$ represent the nuclear norm and entry-wise ℓ_1- norm of a matrix, respectively. However, the problem (2) is nonlinear due to the fact that $\tau \in GL(3)$, and directly minimizing this objective function is expensive. It was shown in [32] that one can linearize the constraint and iteratively estimate a one-step update $\Delta\tau$ by solving

$$\min_{A,E,\tau} \|A\|_* + \lambda \|E\|_1 \quad \text{subj. to} \quad I \circ \tau_k + \nabla I \Delta\tau = A + E. \tag{3}$$

This optimization program then can be solved by algorithms similar to Robust PCA solvers. Figure 1 illustrates the results of applying (3) to some representative low-rank texture regions.

Next, we more rigorously define the clustering problem for segmentation of TILT complexes in natural images:

[1] Without loss of generality, we can assume A and E are square matrices.

Figure 2: Results of the proposed algorithm on two challenging examples in the presence of irregular 3D shapes, vegetation occlusion, and transparent glass surfaces. **Left:** Original images. **Middle Left:** TILT feature detection with local camera frames superimposed (the green arrows indicate surface normals). **Middle Right:** Clusters of TILT complexes in color. The nodes not in the colored clusters are pruned out. **Right:** Fitting higher-level TILT models to the complexes.

Problem 1 (Multiscale TILT Clustering (MTC)) *Given a natural image, the MTC problem seeks solutions to the following three closely related subproblems:*

1. *Obtain a set of TILT features: $I_1, \cdots, I_n$. Each TILT feature is decomposed to $I_k = (A_k, E_k, \tau_k)$, where τ_k represents the homography transformation from the 3D position of the texture pattern in space to the camera, E_k is the sparse pixel corruption matrix, and A_k is the low-rank texture representation.*

2. *The n TILT features form a 2D adjacency graph $G = (V, E)$ called the* TILT adjacency graph (TAG)*, where $V = \{I_1, \cdots, I_n\}$ represents the list of n nodes, and an edge e_{ij} is present, i.e., $e_{ij} \in E$, if two features I_i and I_j are close to each other in the image space. Define a cost function $f(e_{ij})$ associated to the edge e_{ij} that measures the dissimilarity of the two adjacent TILT features in terms of their low-rank components (A_i, A_j) and transformations (τ_i, τ_j).*

3. *Seek an efficient clustering algorithm capable of partitioning the connecting TILT features in the TAG to subgraphs called* TILT complexes*, each of which represents a unique planar structure in space and is occupied by several TILT features within the complex. Furthermore, two adjacent TILT complexes necessarily represent different 3D shapes or different surface texture patterns in space.*

3. Geometric Image Segmentation via MTC

In this section, we discuss in details the design of the MTC algorithm. First, we discuss the detection of TILT features in multiple scales in Section 3.1. Then we discuss how to select optimal TILT scales in an adjacency graph in Section 3.2. Finally, we present an effective algorithm to partition the TILT adjacency graph into subgraphs called TILT complexes in Section 3.3.

3.1. Multiscale TILT Detection

Given a natural image such as the one shown in Figure 3 Left, we first need to partition the image into local patches where the TILT representation is calculated. A popular approach to group local homogeneous texture regions is to use superpixels [23, 8]. In this paper, we choose a public code Quick Shift [28] to pre-segment the image into superpixels due to its fast speed compared to the other methods, as shown in Figure 3 Middle.

After superpixel extraction, each superpixel can be fitted by a bounding box, and the TILT algorithm [32] is readily applied to the bounding box as the initial position of a potential TILT feature. However, in practice, we have observed that directly applying TILT to superpixels may not always yield good representation, even when the superpix-

Figure 3: An example of multiscale TILT feature extraction. **Left**: Original. **Middle**: Superpixels rendered using mean colors. **Right**: Multiscale TILT features. A local frame is attached to each TILT feature to illustrate its 3D orientation with the green arrows indicating the surface normals. The algorithm recovers the correct 3D orientation of the pattern at scale 3 and 4, while the results from the smaller scales are not as accurate.

els represent salient geometric structures in space. One major reason is that often the dimension of a superpixel could be rather small, while the underlying algorithm of Robust PCA that underpins the TILT algorithm requires the input matrix to have a sufficient size for the algorithm to be effective.[2]

Motivated by the solutions of many existing invariant features, we address the above problem by adopting a multiscale scheme. First, we enforce that the first-level bounding box that centers at a superpixel must be at least 50×50 pixels. Second, we create larger bounding boxes that are centered at the same location but whose sizes are 1.2, 1.4, and 1.6 times bigger than the first bounding box, respectively. Then each bounding box for the i-th superpixel at scale j is processed by TILT as $I_i^j = (A_i^j, E_i^j, \tau_i^j)$. As shown in the right plot of Figure 3, TILT estimation at multiple scales better captures repetitive 2D patterns and their homography transforms than at the original superpixel scale.

Before we proceed to select the most consistent TILT features in multiple scales in the next section, we notice that certain regions in an image may represent homogeneous color patches (such as sky and water) or noisy high-rank patches (such as trees, grass, and pedestrians). As a result, the estimation of TILT at those regions may be noisy and not consistent with any meaningful geometric structure in space, as shown in Figure 4. Therefore, these regions should be first excluded from the subsequent MTC calculation. Using the TILT decomposition $I = (A, E, \tau)$, this task can be easily achieved by checking the estimated rank of the low-rank component A.

[2]In fact, the Robust PCA theory that guarantees the exact recovery of the low-rank and sparse components of a matrix holds asymptotically only when the size of the matrix grows large (i.e., towards infinity).

Figure 4: Example of a non-geometric vegetation image in which multiscale TILT features are not consistent.

To do so, we first define the *canonical rank* of an image.

Definition 2 (Canonical Rank) *Given an image patch I and its TILT components $I = (A, E, \tau)$, its canonical rank $\rho(I)$ is defined by thresholding the energy of its low-rank component A in* singular value decomposition*:*

$$\rho(I) \doteq \arg\min_k \frac{\sum_{i=1}^k \sigma_i^2(A)}{\|A\|_F^2} > 0.999, \qquad (4)$$

where σ_i is the i-th singular value of A: $\sigma_1 \geq \sigma_2 \geq \cdots$.

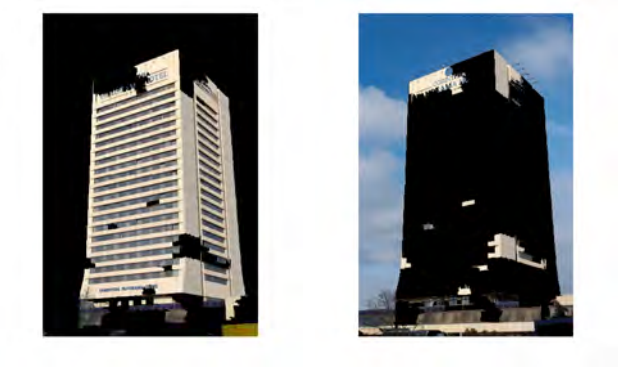
In Definition 2, the choice of the energy threshold at 99.9% is empirical. We have found that this value works well to partition an image into a *geometric layer* and a *non-geometric layer*. More specifically, for an image region I that contains a superpixel, if its canonical rank $\rho(I)$ at any scale is smaller than a preset threshold α_1, then the superpixel will be designated as a homogeneous color region. Similarly, if $\rho(I)$ is greater than another preset threshold α_2 at any scale, then the superpixel will be designated as a noisy region. Color regions and noisy regions typically do not represent geometric structures in space. Hence, we merge these regions to a non-geometric layer. Conversely, if $\alpha_1 \leq \rho(I) \leq \alpha_2$ at all scales, then the superpixel is a low-rank region and belongs to a geometric layer.

Figure 5 shows an example of partitioning an image into the two layers. More examples are shown in Section 4.

3.2. TILT Adjacency Graph

In this section, we assume N superpixels in the geometric layer have been fitted with TILT features at multiple scales (e.g., four as we specified above). The task is to build a 2D adjacency graph G to establish their spatial and texture similarities, which is called a *TILT adjacency graph* (TAG). We will also apply a *Markov random field* (MRF) model on the TAG to select an optimal TILT representation of each superpixel among the multiple scales such that the 3D orientations of neighboring TILT features are consistent. Figure 6 illustrates an example of building the TAG.

First, a TILT adjacency graph (TAG) is defined as $G = (V, E)$, where $V = \{I_1, I_2, \ldots, I_N\}$ is the set of nodes that represent the N superpixels in the geometric layer, and $E = \{e_{ij}\}$ is the set of edges that connect two nodes I_i and I_j if the two superpixels share a common boundary in the image.



(a) Collection of low-rank regions (b) Collection of color/noisy regions

Figure 5: Partition of the image in Figure 3 into (a) the geometric layer and (b) the non-geometric layer.

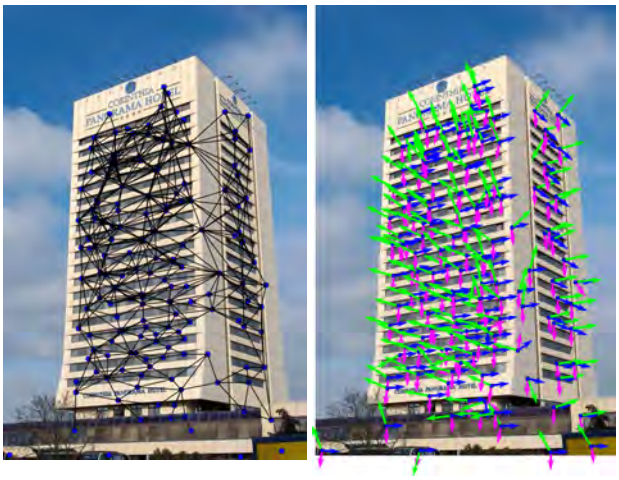


Figure 6: **Left**: The TAG of the image in Figure 3. **Right:** Selection of consistent TILT representation in multiple scales.

Second, based on the estimated TAG, we want to determine the optimal TILT scale from the multiscale representation such that the 3D orientation of the connected TILT features in the TAG are consistent. In this paper, we have chosen four scales at each superpixel to represent its TILT features in Section 3.1. Therefore, the orientation of a superpixel I_i can be represented by four normal vectors $(\boldsymbol{n}_i^1, \boldsymbol{n}_i^2, \boldsymbol{n}_i^3, \boldsymbol{n}_i^4)$.[3] Furthermore, two normal vectors connected in the TAG define a potential function for the MRF:

$$V(\boldsymbol{n}_i, \boldsymbol{n}_j) = \arccos\left(\frac{\boldsymbol{n}_i{}^T \boldsymbol{n}_j}{\|\boldsymbol{n}_i\|_2 \|\boldsymbol{n}_j\|_2}\right). \quad (5)$$

The intuition behind potential function (5) is that a superpixel most likely has the same normal vector as its adjacent superpixels.

Given the potential function and the TAG, the distribution of the candidate TILT features for the combination $X = \{\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n}_N\}$ on the MRF is defined as:

$$P(X) = \frac{1}{Z} \exp(-\sum_{e_{ij} \in E} V(\boldsymbol{n}_i, \boldsymbol{n}_j)), \quad (6)$$

where Z is the normalization value. Finally, we seek the configuration $X^* = \{\boldsymbol{n}_1^*, \boldsymbol{n}_2^*, \ldots, \boldsymbol{n}_N^*\}$ that maximizes the above distribution function:

$$X^* = \arg\max_X P(X). \quad (7)$$

This optimization problem is, in general, NP-hard [3]. In the literature, there exist two classical methods to deal with this problem, which are *iterated conditional modes* [2] and *simulated annealing* based on Gibbs sampling [9]. In our experiment, we have found that both solutions can provide reasonable results for the most likely configuration. Since MRF optimization is not the main focus of this paper, we simply choose the Gibbs sampling method in our algorithm.

Finally, we note that enforcing the TAG and MRF may still group TILT features from structures with different texture patterns if they share similar 3D orientations. Some examples are shown in Section 4. Therefore, we are motivated to further partition the TAG based on the texture similarity of the TILT features. More specifically, we use the well-known Gabor filters [17] and the χ^2-distance [24] to measure the similarity of two texture regions under TILT transform. A 2D mother Gabor wavelet g at coordinates (x, y) is given by:

$$g_{\sigma,\lambda}(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2} + i\frac{2\pi x}{\lambda}\right) \in \mathbb{C}, \quad (8)$$

where σ is the Gaussian localization parameter and λ is the wavelength of the sinusoidal factor. In this paper, we choose a family of 12 self-similar Gabor wavelets derived from $g_{\sigma,\lambda}$ in [17], which contains 3 scales and 4 orientations.

In an TAG, the response of a TILT feature $I^* = (A, E, \tau)$ whose optimal scale is selected by the MRF (7) to a Gabor wavelet function $g^{(i)}$ is defined by the convolution

$$F^{(i)} = \|A * g^{(i)}\| \subset \mathbb{R}^2. \quad (9)$$

The pixel distribution in the convoluted image $F^{(i)}$ can be represented by a normalized histogram vector. Subsequently, the texture similarity $D(I_1, I_2)$ of two TILT features I_1 and I_2 can be calculated by the χ^2-distances of their Gabor histogram vectors over all the wavelet filters [24]. Using the texture similarity metric D, one can choose a quite conservative threshold α_3. If two adjacent TILT features satisfy $D(I_i, I_j) > \alpha_3$, their edge e_{ij} will be removed from the TAG.

[3] A normal vector $\boldsymbol{n}_i^j$ can be recovered from the decomposition of its homography transformation τ_i^j [11, 16].

3.3. Building TILT Complexes

Given the estimated TILT features in the TAG, in this section, we further partition the TAG into subgraphs, each of which represents a global planar structure in space. Subsequently, a larger TILT representation of each complex can be fitted that contains all the TILT features in the subgraph, and hence provides a more global representation of the urban structures. An example is shown in Figure 7.

Figure 7: **Left:** Partitioning the TAG in Figure 6 into two complexes connected by red and cyan edges (in color). Outlying TILT features that are not connected to the two subgraphs are also excluded. **Right:** The two TILT complexes provide a higher-level global geometric model.

First, we observe that if two adjacent superpixels I_i and I_j belong to the same facade, they often share similar texture patterns *and* orientations. As the cue of texture similarity has been utilized in the construction of the TAG above, a naive way to take advantage of the other geometric cue is to directly compare the similarity of their homographies (τ_i, τ_j) from their TILT representations. However, we have found that in practice, especially in urban images, two planar structures such as building facades might share similar textures and orientations in space, but they could represent two complete different 3D surfaces with different depths in space. Such regions that are similar only based on their local TILT representations should not be merged and treated as a single planar structure.

To mitigate this problem and inspired by the work in [32], we propose to introduce a verification step that hypothetically merge I_i and I_j as a new image $I_{ij} \doteq [I_i, I_j]$[4], and again solve its TILT representation as:

$$\min_{A', E', \tau_{ij}} \|A'\|_* + \lambda \|E'\|_1 \quad \text{subj. to} \quad I_{ij} \circ \tau_{ij} = A' + E'. \tag{10}$$

[4]By an abuse of notation, I_{ij} is the minimal bounding-box region that contains both I_i and I_j and other pixels in between.

We define another cost function $f(\boldsymbol{n}_i, \boldsymbol{n}_j, \boldsymbol{n}_{ij})$ on the TAG associated to the edge e_{ij} that measures the dissimilarity of the two adjacent TILT features in terms of their orientations $(\boldsymbol{n}_i, \boldsymbol{n}_j, \boldsymbol{n}_{ij})$, which are calculated from $(\tau_i, \tau_j, \tau_{ij})$ as

$$f(\boldsymbol{n}_i, \boldsymbol{n}_j, \boldsymbol{n}_{ij}) = \exp\left(-\frac{\alpha_4}{\max(V(\boldsymbol{n}_i, \boldsymbol{n}_{ij}), V(\boldsymbol{n}_j, \boldsymbol{n}_{ij}))^2}\right), \tag{11}$$

where $0 \leq f(\cdot) < 1$ and α_4 is a user-defined parameter. When I_i and I_j are with the same facade, ideally $\boldsymbol{n}_i = \boldsymbol{n}_j = \boldsymbol{n}_{ij}$ so that $f(\cdot) = 0$. Therefore, the problem of clustering TILT features into TILT complexes becomes a graph partitioning problem on the TAG.

For two main reasons, instead of using a standard graph-cut algorithm such as [26], we use the recently proposed *dissimilarity-based sparse representation selection* (DSRS) algorithm [7] for graph partitioning. First, some of the nodes in the graph correspond to outlying features since the corresponding superpixels contain different regions, such as two different facades or a facade occluded by trees. Second, the number of clusters is not known a priori. Such problems cannot be reliably handled by traditional graph partitioning techniques such as the Normalized Cut algorithm [26]. On the other hand, DSRS algorithm can robustly cluster the graph for a large range of its single regularization parameter and can also reject outliers [7]. However, the algorithm requires to have dissimilarities between all pairs of connected nodes. Thus, to take advantage of the DSRS algorithm, we define the dissimilarity $f(\cdot)$ between any two nodes as the total dissimilarity on the shortest path between the connected nodes on the TAG. The output of the algorithm finds clustering of the nodes, while the outliers as whose subgraphs with very small sizes are detected and rejected.

4. Experiment

For our experiment, we use the Pankrac dataset [6], which consists of 82 images of 30 urban buildings. For the user defined parameters in the MTC algorithm, we set $\alpha_1 = 1$, $\alpha_2 = 13$, $\alpha_3 = 1.5$, and $\alpha_4 = 0.2$.

Figure 8 highlights some representative examples of geometric image segmentation using our algorithm. These examples demonstrate that our algorithm is capable of finding dominant geometric structures in a wide variety of conditions:

1. In all the results shown in the paper, the image regions with no TILT feature attached belong to the estimated non-geometric layer. Utilizing the multiscale TILT detection and the canonical rank condition, our algorithm is able to accurately partition an image into geometric and non-geometric layers.

2. We observe that the MRF model is very effective in selecting consistent local TILT features at optimal scales, even when the planar structures have large non-Lambertian surfaces (i.e., glass) and/or large perspective distortion.

3. As shown in the first four examples, surfaces with similar texture patterns may have very different 3D orientations and depths. Our proposed method using DSRS successfully clusters the TILT features into TILT complexes. The more global TILT complex models accurately describe the overall 3D shape of the large buildings in space.

4. The DSRS algorithm also effectively prunes out outlying TILT features that are not from the dominant planar structures.

In terms of the speed, the complexity of the full pipeline is clearly dominated by the calculation of TILT representation at multiple scales. The reader is referred to [4, 32, 21] for fast TILT solvers. Our algorithm that builds global models adds to the complexity by requiring a verification step (10). However, this step can be ignored if one only cares about the detection of consistent local TILT features.

Finally, we discuss a few examples where our algorithm returns inaccurate geometric segmentation in Figure 9:

1. *Non-Lambertian surfaces*. Our method is not completely immune from the effect of non-Lambertian surfaces, which lead to inconsistent TILT features. In the first example in Figure 9, the windows reflect the texture of the sky and clouds, and they are sometimes transparent as well.

2. *Similar local texture and orientation*. In the second example, the left TILT complex contains facades from two adjacent buildings, which share similar textures and orientations. Nevertheless, for the building on the right, its planar structure is correctly recovered.

3. *Lack of texture*. If a facade is covered primarily by homogenous color, our algorithm will exclude its TILT features. In the third example, the algorithm still correctly detects the TILT complex on the top that has richer texture appearance.

Figure 8: Representative examples of geometric image segmentation. **Left:** Original image. **Middle Left:** TILT detection. **Middle Right:** TILT complexes. **Right:** Fitting higher-level TILT representation to TILT complexes.

5. Conclusion

Compared to traditional image features, global geometric features such as TILT have shown attractive attributes that may pertain to several high-level vision applications. However, they have not been widely used in the past mainly due to lack of effective algorithms to detect low-rank image regions from natural images. This paper addresses this gap via a novel multiscale TILT clustering algorithm as a means of geometric segmentation. The algorithm can be used as a fundamental image feature detection

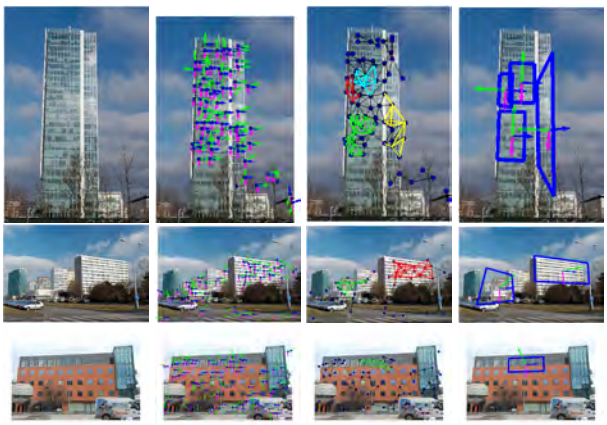


Figure 9: Some inaccurate geometric segmentation results.

method that complements the existing invariant feature detection algorithms, especially for urban images where symmetric man-made structures abound.

Acknowledgment

The authors are grateful to Dr. Tony Xu Han at the University of Missouri, Nikhil Naikal at the University of California, Berkeley, and Dr. Zihan Zhou at the Pennsylvania State University for their useful suggestions.

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008. 1

[2] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):pp. 259–302, 1986. 5

[3] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 1998. 5

[4] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis. *J. ACM*, 58(1):1–37, 2011. 1, 2, 7

[5] A. Cohen, C. Zach, S. Sinha, and M. Pollefeys. Discovering and exploiting 3d symmetries in structure from motion. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2012. 1

[6] P. Doubek, J. Matas, M. Perdoch, and O. Chum. Image matching and retrieval by repetitive patterns. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3195–3198, 2010. 6

[7] E. Elhamifar, G. Sapiro, and R. Vidal. Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. In *NIPS*, 2012. 2, 6

[8] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *International Journal on Computer Vision*, September 2004. 3

[9] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984. 5

[10] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *Proceedings of the IEEE International Conference on Computer Vision*, 2009. 1

[11] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, 2000. 5

[12] W. Hong, A. Yang, and Y. Ma. On symmetry and multiple view geometry: Structure, pose and calibration from a single image. *International Journal on Computer Vision*, 60:241–265, 2004. 1

[13] K. Koeser, C. Zach, and M. Pollefeys. Dense 3D reconstruction of symmetric scenes from a single image. In *Annual Symposium of the German Association for Pattern Recognition (DAGM)*, 2011. 1

[14] Y. Liu, H. Hel-Or, C. Kaplan, and L. van Gool. Computational symmetry in computer vision and computer graphics. *FTCGV*, 5:1–191, 2010. 1

[15] D. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*, 1999. 1

[16] Y. Ma, J. Košecká, S. Soatto, and S. Sastry. *An Invitation to 3-D Vision, From Images to Models*. Springer-Verlag, New York, 2004. 5

[17] B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996. 5

[18] J. Matas, O. Chum, M. Urba, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *British Machine Vision Conference*, pages 384–396, 2002. 1

[19] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal on Computer Vision*, 65(1–2):43–72, 2005. 1

[20] B. Mičušík and J. Košecká. Piecewise planar city 3D modeling from street view panaramic sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, 2009. 1

[21] H. Mobahi, Z. Zhou, A. Yang, and Y. Ma. Holistic 3D reconstruction of urban structures from low-rank textures. In *ICCV Workshop on 3D Representation and Recognition*, 2011. 2, 7

[22] G. Mori. Guiding model search using segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2005. 1

[23] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: combining segmentation and recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2004. 3

[24] Y. Rubner, J. Puzicha, C. Tomasi, and J. Buhmann. Empirical evaluation of disimilarity measures for color and texture. *Computer Vision and Image Understanding*, 84:25–43, 2001. 5

[25] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Efficient structured prediction for 3D indoor scene understanding. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2012. 1

[26] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 731–737, 1997. 6

[27] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011. 1

[28] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. *10th European Conference on Computer Vision*, pages 705–718, 2008. 3

[29] A. Yang, S. Rao, K. Huang, W. Hong, and Y. Ma. Symmetry-based 3-d reconstruction from perspective images. *Computer Vision and Image Understanding*, 99:210–240, 2005. 1

[30] Z. Zhang, Y. Matsushita, and Y. Ma. Camera calibration with lens distortion from low-rank textures. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2011. 2

[31] P. Zhao and L. Quan. Translation symmetry detection in a fronto-parallel view. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2011. 2

[32] Z. Zhou, X. Liang, A. Ganesh, and Y. Ma. TILT: Transform invariant low-rank textures. In *Proceedings of Asian Conference on Computer Vision*, 2010. 1, 2, 3, 6, 7

[33] S. Zhu and D. Mumford. A stochastic grammar of images. *FTCGV*, 2006. 1