Multiscale TILT Feature Detection with Application to Geometric Image Segmentation [*]

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

Motivated by the theory of low-rank matrix representation, a new type of invariant image feature, called transform-invariant low-rank texture (TILT), has been recently proposed. However, the applicability of TILT features in computer vision has been severely limited by two major problems. First, TILT feature representation is based on the assumption that the given image contains only one dominant low-rank region, which typically does not hold in natural images. Second, when multiple low-rank regions are present, the existing TILT detection methods either randomly sample the image or apply to fixed grid coordinates, both of which cannot guarantee good recovery of salient low-rank image features. In this paper, we propose a novel algorithm to address these two important issues. First, utilizing superpixels and the concept of canonical rank derived from TILT, we introduce a method to segment natural images into a geometric layer and a non-geometric layer. Second, we apply a Markov random field model to a multiscale low-rank representation of the image geometric layer, and obtain an effective algorithm to detect TILT features. Finally, we present an application of the multiscale TILT detection algorithm to the classical problem of building facade segmentation. Extensive experiments are conducted on the Pankrac building database to demonstrate the efficacy of the algorithms.

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 [16, 19, 20, 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 [32, 15, 5]. Using the virtual views of symmetric patterns, their 3D orientation can be readily estimated from just a single image [13, 14]. Another type of geometric features used in 3D modeling are homogeneous color regions such as superpixels [23] whose orientation under perspective projection is consistent with that of some global planar structures in space [21, 27]. Finally, in object recognition and segmentation, various types of object part-based regions that contain rich semantic information have been proposed [36, 11, 30].

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) [35]. 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

[*] This work was supported in part by DARPA FA8650-11-1-7153, ARO 63092-MA-II, and ONR N00014-09-1-0230.

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 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.

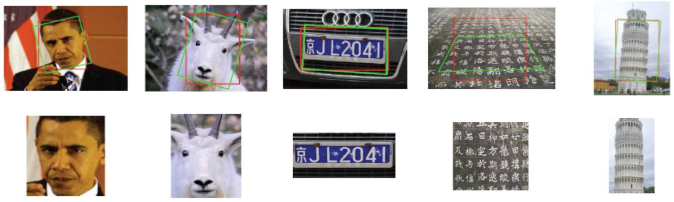


Figure 1. Examples of manually labeled image patterns that are extracted as TILT features. **Top:** Initialization of the feature locations as the red bounding boxes, and the final orientation of the feature as the green bounding boxes. The TILT features compensate the perspective distortion. **Bottom:**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

We are aware of three applications where the use of TILT features has been considered: 3D reconstruction of building facades [22], symmetry detection [34], and camera calibration [33]. 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, despite attractive attributes of TILT, it has not been widely adopted in other vision application where the use of invariant features would be preferred. We believe this is mainly due to two reasons. First, TILT feature representation is based on the assumption that the given image contains only one dominant low-rank region, which typically does not hold in natural images. Second, when multiple low-rank regions are present, the existing TILT detection methods either randomly sample the image or apply to fixed grid coordinates, both of which cannot guarantee good recovery of salient low-rank image features.

1.1. Contributions

In this paper, we propose a novel algorithm called *multi-scale TILT detection* (MTD) to address the above two critical issues that have handicapped the use of TILT features in computer vision applications. First, utilizing superpixels and the concept of canonical rank derived from TILT, we introduce a method to segment natural images into a geometric layer and a non-geometric layer. Second, we apply a Markov random field model to a multiscale low-rank representation of the image geometric layer, and obtain an effective algorithm to detect TILT features. To this 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.

We believe the new TILT detection algorithm can be readily employed by higher-level algorithms in object recognition, image retrieval, and 3D reconstruction. In this paper, we present an example to apply the algorithm to the classical problem of modeling 3D planar structures. 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, we employ a robust clustering algorithm to cluster the graph into multiple groups while rejecting the outlying nodes. Each of the groups represents a dominant planar structure, e.g., a building facade.

Note that, when MTD is applied to the application of finding building facades, the result bears resemblance to a category of urban scene reconstruction algorithms based on detecting image texture symmetry such as [25]. Nevertheless, the focus of most symmetry-based facade modeling algorithms including [25] is on finding 2D deformable lattice structures, and the implementation is typically based on some existing salient features such as SIFT features. In this paper, our main focus is on the detection of more robust TILT features from natural images. Arguably, the TILT features can be also used as the basic "building block" to construct 2D lattice structures. More importantly, since TILT features are more robust in handling camera perspective distortion, illumination change, and pixel occlusion, more complex facade structures can be successfully recovered by the MTD algorithm, as some examples shown in our experiments.

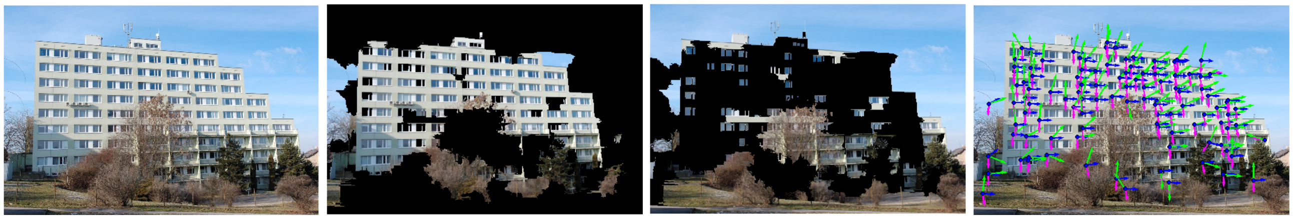
2. Problem Formulation

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 $I = (A, E, \tau)$,

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



(a) Original image (b) Geometric layer (c) Non-geometric layer (d) TILT feature detection

Figure 2. Results of the proposed algorithm on a challenging example in the presence of perspective distortion, vegetation occlusion, and transparent glass surfaces. TILT feature detection results are illustrated by the superimposed local frames (the green arrows indicate surface normals).

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 [35] that one can linearize the constraint and iteratively estimate a one-step update $\Delta\tau$ by solving

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

This optimization program then can be solved by algorithms similar to Robust PCA solvers. Figure 1 illustrates the results of applying (2) to some representative low-rank texture regions.

Next, we more rigorously define the multiscale TILT detection problem:

Problem 1 (Multiscale TILT Detection (MTD)) *Given a natural image, the MTD problem seeks solutions to obtain a set of TILT features from unique image regions: $I_1, \cdots, I_n$. 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.*

3. Multiscale TILT Detection

3.1. Multiscale Low-Rank Analysis

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 [24, 9], as shown in Figure 3 Middle.

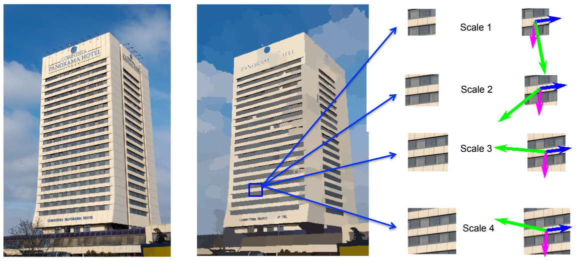


Figure 3. An example of multiscale TILT feature extraction. **Left**: Original. **Middle**: Superpixels rendered using mean colors. **Right**: Multiscale TILT features. A local frame is attached to each TILT feature to illustrate its 3D orientation with the green arrows indicating the surface normals. 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.

After superpixel extraction, each superpixel can be fitted by a bounding box, and the TILT algorithm [35] is readily applied to the bounding box as the initial position of a potential TILT feature. However, in practice, directly applying TILT to superpixels may not always yield good representation, even when the superpixels 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]

We address the above problem by adopting a multiscale scheme similar to other invariant feature detection algorithms such as SIFT. More specifically, at each superpixel, we consider bounding boxes of increasing sizes. In other words, we first consider a bounding box of size $w \times w$ centered at each superpixel. Next, we increase the size of the

[2] In fact, 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.

original bounding box by ratios $r_0 = 1 < r_1 < r_2 < \cdots < r_{L-1}$, corresponding to L different scales. Then, for the i-th superpixel, its bounding box 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.

In the next section, for neighboring superpixels, we will select their most consistent TILT features at multiple scales. Before that, it is important to 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 MTD 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.

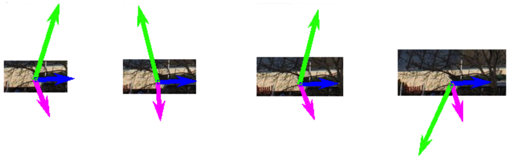


Figure 4. Example of a non-geometric vegetation image in which multiscale TILT features are not consistent.

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

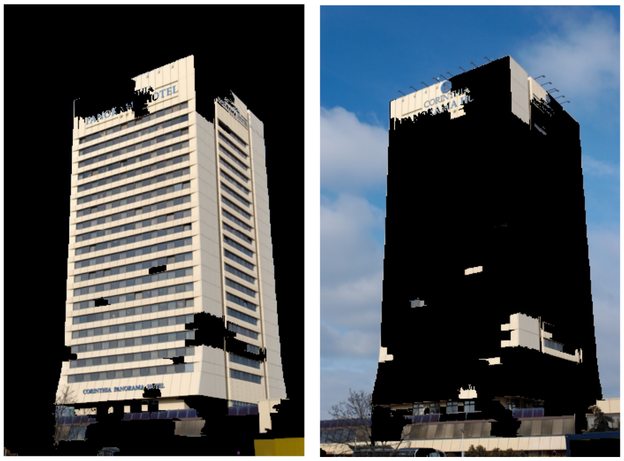
Definition 2 (Canonical Rank) *Given an image patch I and its TILT components $I = (A, E, \tau)$, its canonical rank $\rho(I)$ is defined by thresholding the energy of its low-rank component A in* singular value decomposition*:*

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

where σ_i is the i-th singular value of A: $\sigma_1 \geq \sigma_2 \geq \cdots$, and γ is a predetermined fidelity threshold.

We have found that $\rho(I)$ provides a good criterion 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 5.



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

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

3.2. TILT Detection on 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 shown in Figure 3). 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.

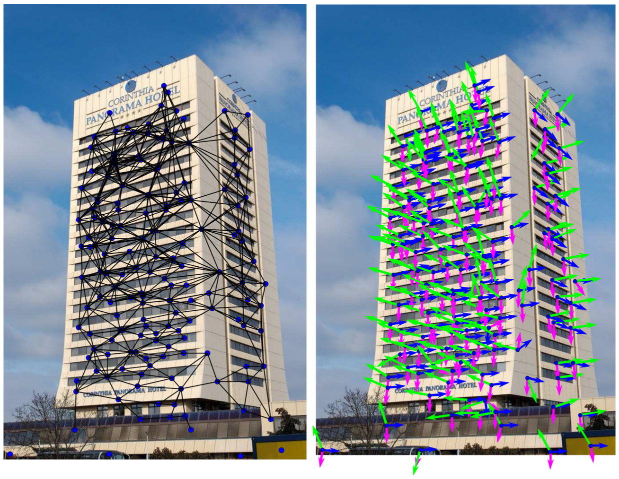


Figure 6. **Left**: The TAG of the image in Figure 3. **Right:** Selection of consistent TILT representation in multiple scales.

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.

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:

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

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 $X = \{\boldsymbol{n}_1, \boldsymbol{n}_2, \dots, \boldsymbol{n}_N\}$ on the MRF is defined as:

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

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

$$X^* = \arg\max_X P(X). \tag{7}$$

Since solving the above optimization program is, in general, NP-hard [3], in practice we can use two classical methods, which are *iterated conditional modes* [2] and *simulated annealing* based on Gibbs sampling [10]. 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.

The MTD algorithm is summarized as follows.

4. Application: Modeling Planar Structures

In this section, we demonstrate an illustrative application of TILT features to modeling planar structures. The basic idea is that, given the TILT features in the TAG, we partition the TAG into subgraphs, each of which represents a global planar structure in space. Subsequently, a larger TILT representation of each group can be fitted that contains all the TILT features in the subgraph, called a *TILT complex*. A TILT complex provides a geometric representation of even larger, more global urban structures in natural images. An example is shown in Figure 7.

Algorithm 1: Multiscale TILT Detection (MTD)

```
1: Partition the input image into superpixels $I_1, \dots, I_N$.
2: Compute multiscale TILT representation
   $I_i^j = (A_i^j, E_i^j, \tau_i^j)$
3: Partition all superpixels into a geometric layer and a
   non-geometric layer based on canonical rank (4).
4: Build TILT adjacency graph $G$.
5: Determine optimal TILT scale $X^*$ by maximizing the
   MRF distribution (7).
Output: TILT detection on the geometric layer.
```

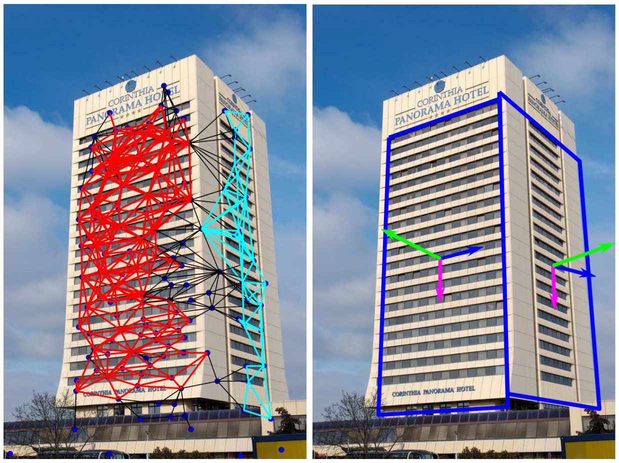


Figure 7. **Left:** Partitioning the TAG in Figure 6 into two complexes connected by red and cyan edges (in color). Outlying TILT features that are not connected to the two subgraphs are also excluded. **Right:** The two TILT complexes provide a higher-level global geometric model.

First, we 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 5. 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 [18] and the χ^2-distance [26] to measure the similarity of two texture regions under TILT transform. A 2D mother Gabor wavelet g at coordinates (x, y) is given by:

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

where σ is the Gaussian localization parameter and λ is the wavelength of the sinusoidal factor.

In a 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

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

[3] A normal vector $\boldsymbol{n}_i^j$ can be recovered from the decomposition of its homography transformation τ_i^j [12, 17].

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 [26]. If the texture similarity distance $D(I_i, I_j)$ of two TILT features is too large, their edge e_{ij} will be removed from the TAG.

Second, 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 [35], 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 combined TILT representation as:

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

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

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

where $0 \leq f(\cdot) < 1$ and α 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 [29], we employ the recently proposed *dissimilarity-based sparse modeling representative selection* (DSMRS) algorithm [8] 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 [29]. On the other hand, DSMRS algorithm can robustly cluster the graph for a large range of its single regularization parameter and can also reject outliers [8]. However, the algorithm requires to have dissimilarities between all pairs of connected nodes. Thus, to take advantage of the DSMRS 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.

5. Implementation and Experiment

In the implementation of the MTD algorithm, we choose a public code Quick Shift [31] to pre-segment the image into superpixels due to its fast speed compared to other existing methods like mean shift [6] and Medoidshifts [28]. We choose the initial window size around each superpixel as $w = 50$ pixels and consider $L = 4$ scales with $r_1 = 1.2$, $r_2 = 1.4$ and $r_3 = 1.6$. We choose the rank threshold in equation (4) as $\gamma = 0.999$, which yields good empirical results. To solve the MRF problem in (7), we use the Gibbs sampling algorithm. However, we have observed that the iterated conditional modes algorithm also provides equally good results as the Gibbs sampling method. For the user defined parameters in the MTD algorithm, we set $\alpha_1 = 1$, $\alpha_2 = 13$. We have observed that the MTD algorithm is not very sensitive to the change of these values due to the robustness of the method.

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, 35, 22] for fast TILT solvers.

5.1. Multiscale TILT Detection

The two experiments in this and next section are based on the Pankrac dataset [7], which consists of 82 images of 30 urban buildings.

First, we apply the MTD algorithm on the Pankrac dataset. The TILT feature detection results are shown in the first two columns of Figure 8. In all the results, the image regions with no TILT feature attached belong to the estimated non-geometric layer.

It is worth noting that the first two images in Figure 8 contain significant portions of non-Lambertian surfaces and complex sky texture. The MTD algorithm is able to segment the sky region into the non-geometric layer, and accurately recover the 3D orientation of the TILT features on the building glass surface. In the next three examples, MTD successfully recovers the TILT features on more complex

[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.

3D shapes, many of which do not satisfy any 2D symmetry models.

Finally, we draw the following conclusions:

1. Utilizing the canonical rank condition, our algorithm is able to accurately partition an image into the geometric and non-geometric layers.

2. 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.

5.2. Building Facade Detection

In this experiment, we demonstrate the application of modeling urban building facades using TILT complexes. The results from the same Pankrac images are shown in the last two columns of Figure 8.

We observe that the DSMRS algorithm is capable of finding dominant geometric structures in a wide variety of conditions. In particular, in the first four examples, surfaces with different orientation or different texture patterns are correctly segmented. In the last example, surfaces with similar texture but different depths are also segmented. The more global TILT complex models accurately describe the overall 3D shapes of the large buildings in space. The algorithm also effectively prunes out outlying TILT features that are not from the dominant planar structures.

Compared to the existing facade modeling algorithms such as [25], those algorithms likely will fail to group the lattice points when their facade features do not have uniform Lambertian appearance or are not translationally symmetric, such as the first five examples of Figure 8.

6. 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. This paper addresses the detection of TILT features via a novel multiscale clustering algorithm as a means of geometric segmentation. The algorithm can be used as a fundamental image feature detection method that complements the existing invariant feature detection algorithms, especially for urban images where symmetric man-made structures abound.

References

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

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

[3] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *CVPR*, 1998. 5

[4] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis. *J. ACM*, 58(1):1–37, 2011. 1, 3, 6

[5] A. Cohen, C. Zach, S. Sinha, and M. Pollefeys. Discoverying and exploiting 3d symmetries in structure from motion. In *CVPR*, 2012. 1

[6] D. Comanicu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 24:603–619, May 2002. 6

[7] P. Doubek, J. Matas, M. Perdoch, and O. Chum. Image matching and retrieval by repetitive patterns. In *ICPR*, pages 3195–3198, 2010. 6

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

[9] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, September 2004. 3

[10] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *PAMI*, 6(6):721–741, 1984. 5

[11] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009. 1

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

[13] W. Hong, A. Yang, and Y. Ma. On symmetry and multiple view geometry: Structure, pose and calibration from a single image. *IJCV*, 60:241–265, 2004. 1

[14] K. Koeser, C. Zach, and M. Pollefeys. Dense 3D reconstruction of symmetric scenes from a single image. In *DAGM*, 2011. 1

[15] 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

[16] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999. 1

[17] 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

[18] B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *PAMI*, 18(8):837–842, 1996. 5

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

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

[21] B. Mičušík and J. Košecká. Piecewise planar city 3D modeling from street view panaramic sequences. In *ICCV*, 2009. 1

[22] 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, 6

[23] G. Mori. Guiding model search using segmentation. In *ICCV*, 2005. 1

[24] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: combining segmentation and recognition. In *CVPR*, 2004. 3

[25] M. Park, K. Brocklehurst, R. Collins, and Y. Liu. Translation-symmtry-based perceptual grouping with applications to urban scenes. In *ACCV*, 2010. 2, 7

[26] Y. Rubner, J. Puzicha, C. Tomasi, and J. Buhmann. Empirical evaluation of disimilarity measures for color and texture. *CVIU*, 84:25–43, 2001. 5, 6

[27] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Efficient structured prediction for 3D indoor scene understanding. In *CVPR*, 2012. 1

[28] Y. Sheikh, E. Khan, and T. Kanade. Mode-seeking by medoidshifts. In *ICCV*, pages 1–8, 2007. 6

[29] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. In *CVPR*, pages 731–737, 1997. 6

[30] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV*, 2011. 1

[31] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. *ECCV*, pages 705–718, 2008. 6

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

[33] Z. Zhang, Y. Matsushita, and Y. Ma. Camera calibration with lens distortion from low-rank textures. In *CVPR*, 2011. 2

[34] P. Zhao and L. Quan. Translation symmetry detection in a fronto-parallel view. In *CVPR*, 2011. 2

[35] Z. Zhou, X. Liang, A. Ganesh, and Y. Ma. TILT: Transform invariant low-rank textures. In *ACCV*, 2010. 1, 3, 6

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

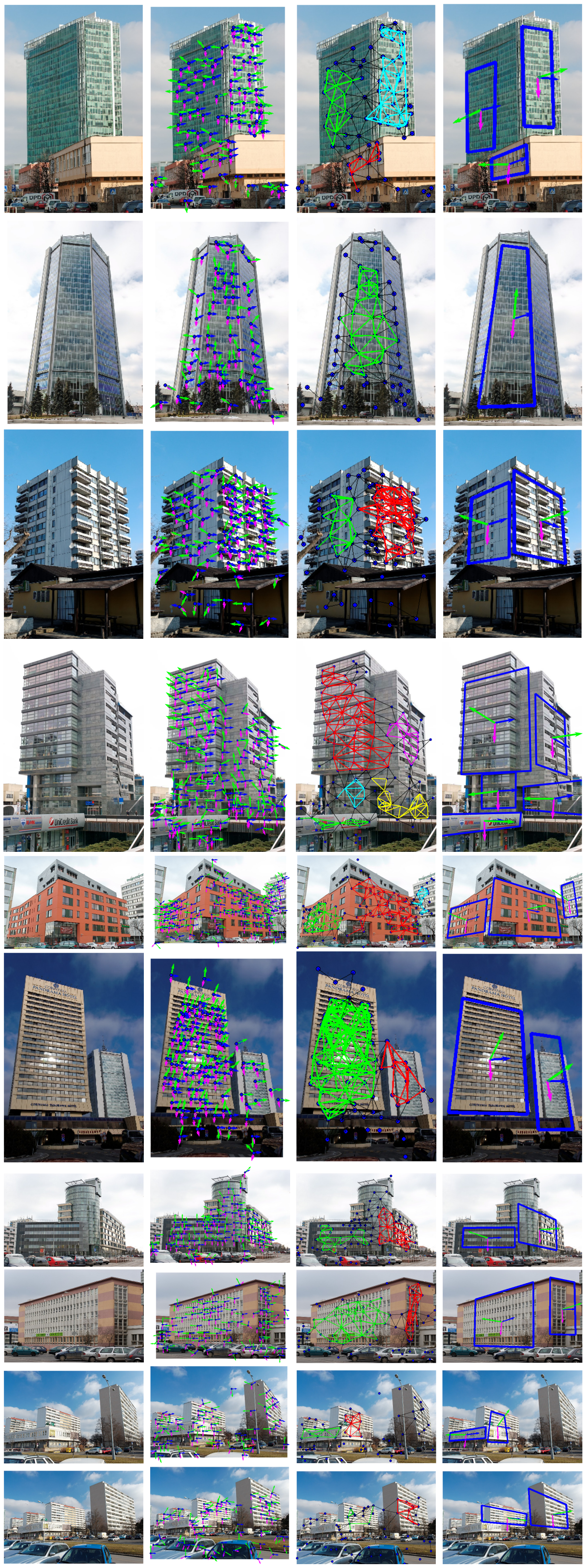


Figure 8. **Left:** Original image. **Middle Left:** Multiscale TILT detection. **Middle Right:** TILT complexes. **Right:** Fitting higher-level TILT representation to TILT complexes.