Facilitating and Exploring Planar Homogeneous Texture for Indoor Scene Understanding

Shahzor Ahmad(✉) and Loong-Fah Cheong

Department of ECE, National University of Singapore, Singapore, Singapore
shahzor.ahmad@gmail.com, eleclf@nus.edu.sg

Abstract. Indoor scenes tend to be abundant with planar homogeneous texture, manifesting as regularly repeating scene elements along a plane. In this work, we propose to exploit such structure to facilitate high-level scene understanding. By robustly fitting a texture projection model to optimal dominant frequency estimates in image patches, we arrive at a projective-invariant method to localize such semantically meaningful regions in multi-planar scenes. The recovered projective parameters also allow an affine-ambiguous rectification in real-world images marred with outliers, room clutter, and photometric severities. Qualitative and quantitative results show our method outperforms existing representative work for both rectification and detection. We then explore the potential of homogeneous texture for two indoor scene understanding tasks. In scenes where vanishing points cannot be reliably detected, or the Manhattan assumption is not satisfied, homogeneous texture detected by the proposed approach provides alternative cues to obtain an indoor scene geometric layout. Second, low-level feature descriptors extracted upon affine rectification of detected texture are found to be not only class-discriminative but also complementary to features without rectification, improving recognition performance on the MIT Indoor67 benchmark.

Keywords: Homogeneous texture · Shape from texture · Planar rectification · Invariant detection · Indoor scene understanding · Geometric layout · Scene classification

1 Introduction

Man-made environments abound with varied manifestation of planar homogeneous texture, i.e., regularly repeating structure or motifs aligned along planes. Figure 1 depicts such "texture" from various indoor scenes in the MIT Indoor67 dataset [1] — repeating objects defining scene content (stacked laundry machines, bookshelves, wine barrels, theater seating, etc.), architectural and

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46475-6_3) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part II, LNCS 9906, pp. 35–51, 2016.
DOI: 10.1007/978-3-319-46475-6_3

Fig. 1. Abundantly present and variedly manifested, homogeneous texture in indoor scenes can serve as useful mid-level features for recognition. Rectification of such texture can mitigate in-class variation arising out of perspective projection.

structural elements (brickwork, frameworks, repeating beams and columns), carpets printed or engraved with uniform patterns, tilings, ceiling fixtures, and even shadows (provided the light source is sufficiently far away, and the blocking objects uniformly spaced)! Such ubiquitous textures must have great potential for favorably influencing high-level scene understanding. Yet, the tools currently at our disposal are woefully inadequate to the purpose of detecting and analyzing textured regions "in the wild", key for realizing the aforementioned potential. In this paper, we examine the technical challenges in detecting these textured regions, develop the machinery necessary to overcome these challenges, and then exploit these textured regions for scene understanding.

Even though invariant texture description and recognition have received regular attention for decades in computer vision, these low-level vision tasks have not been actively pursued as a means to solve high-level vision problems. The reasons are manifold. Firstly, it is difficult to secure a precise definition for texture [2], its optimal representation often necessitating a variety of different mechanisms (such as reaction-diffusion model, grey-level co-occurrence, transform methods, etc.) depending on the circumstances. The same texture can also look significantly different at different scales. When we want to detect and analyze textures in the wild (that is, the textured regions have not been segmented or cropped), the task is complicated by another order of magnitude. Figure 2 illustrates, using some MIT Indoor67 images, the **challenges** involved in localizing such patterns. The texture of interest is often interleaved with other scene content, and such outliers can often occupy large spatial extent — e.g. aisles separating seating sections, beams or arches interfering with repeating columns, visible backdrop through a colonnade, or music stands cluttering patterned woodwork on a concert stage. Photometric severities may be present, such as reflections blocking out under-water pool lanes, low image contrast or varying illumination over a given texture due to insufficient lighting in underground cellars. Finally, texture projected to the image plane inevitably exhibits perspective distortion. Existing region extractors [3] only afford affine invariance, detecting low-level features such as blobs and edges, and cannot localize potentially large patches depicting meaningful texture. In this regard, **our first contribution** is the **projective-invariant detection** of homogeneous textured planar patches in real-world images, as well as their **affine rectification**. In Fig. 2, our approach

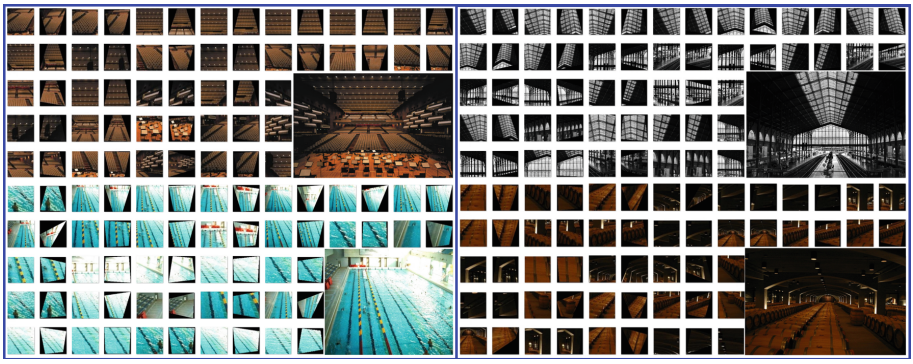

Fig. 2. Detection in the wild: the proposed method can detect and rectify meaningful planar homogeneous texture in indoor scenes, despite outliers with large spatial support, photometric severities and significant perspective distortion. Clockwise: `concert hall`, `train station`, `wine cellar`, `swimming pool`.

is seen to successfully overcome the aforementioned challenges, commonplace in real images. We also present **quantitative evaluations** of our method, outperforming existing work on the tasks of detection and rectification.

Our second contribution is to show how detected homogeneous texture, and their recovered projective parameters, can be used to obtain **indoor geometric layouts** in multi-planar textured scenes. In doing so, we sidestep the error-prone, ill-posed computation of vanishing points in order to establish room orientation, and eschew the simplistic Manhattan or box layout assumption [4]. This also contrasts with existing work [5] that employs machine learning to localize room faces in space and scale.

As seen in Fig. 1, semantically similar image patches can exhibit significant viewpoint differences. Since gradient based low-level image descriptors used in recognition such as SIFT or HOG are not invariant to projective transforms, this can adversely affect classification performance. **Our third contribution** is to demonstrate that plane projective rectification can potentially benefit a recognition pipeline by mitigating this geometric in-class variation. We report improved classification on the MIT Indoor67 [1] benchmark when densely extracted descriptors from affine-rectified texture are included in the image representation, suggesting the feasibility of employing texture cues to achieve rectification in realistic scenes, which, in turn, expectedly improves recognition performance.

2 Related Work

Textured patch detection and rectification are often performed together since by rectifying the perspective effect, the repeating patterns or symmetries are more easily detected. This can be done by exploiting recurring instances of low-level features [6–12], leveraging on different classes of symmetries detected in

the images [13,14], or by rank minimization of the intensity matrix [15]. However, most of these works require restrictive assumptions, e.g. specific symmetries, that the repeating elements form a lattice, that the symmetry type or the repeating element is given, etc. These are serious qualifications in the face of the real-life challenges discussed in the preceding section. Thus, despite the long line of works cited above, there is a paucity of evidence that these methods can work on real images collected in the wild, since they have not been demonstrated on images as rich and complex as say, those found in the benchmark MIT Indoor67, but mainly on limited textures such as building facades, text, or even just pre-segmented or cropped patterns. Different from these approaches, we have adopted a frequency based approach [16] in this paper, as it is capable of describing any generic homogeneous texture (from portholes in laundry machines to shadows — see Fig. 1), not necessarily composed of texels that can be sensed by a given feature detector (lines, blobs, edges, etc.). While the TILT algorithm of [15] also does not involve feature detection, it is, however, applicable to a very limited class of texture — that which upon rectification gives a low-rank matrix. Homogeneity, on the other hand, is a more general assumption.

Shape-From-Texture (SFT). Our work is also related to classical shape-from-texture (SFT) theory — in particular the class of methods that work with planar homogeneous texture [16–18]. However, unlike SFT, our goal is not to recover surface normal, but to perform planar rectification. We therefore re-parameterize the local change in dominant texture frequency [16,19,20] as a function of the plane projective homography instead of the surface slant and tilt. The resulting formulation circumvents the need to define and relate coordinate systems and, more importantly, does not require knowledge of focal length, hence has wider applicability. [21] have previously performed SFT without a calibrated camera, jointly recovering surface normal and focal length, but assume the fronto-parallel appearance of the texture is known a priori. On the other hand, we only make the weak assumption of texture homogeneity. Criminsi and Zisserman [22] also recover vanishing lines from projected homogeneous texture, but their approach involves a computationally expensive search for the direction of maximum variance of a similarity measure, seems to be susceptible to such parameters as the size of image patch to compute the measure over, and has only been demonstrated on cropped texture exhibiting a grid structure.

Scene Layout. Automatic detection of dominant rectangular planar structure has been previously presented in limited, simplistic, and non-cluttered man-made indoor [23] or urban [24] environments. [5] have demonstrated the localization of primary indoor room faces (walls, ceiling, floor) by employing sophisticated machine learning, while [25] have detected depth-ordered planes. However, *all* these approaches assume the scene is aligned with a triplet of principal directions defining the coordinate frame (Manhattan layout), and that these directions can be reliably recovered in a scene. On the other hand, our method detects homogeneous texture to recover geometric layout in multi-planar indoor scenes that do not necessarily conform to the above assumptions.

Indoor Scene Recognition. Since indoor scenes can be well described by the objects and components they contain, their recognition has often been approached through the detection of class-discriminative, mid-level visual features or parts that preserve semantics and spatial information [1]. Automatic part learning from images labeled only with the scene category has received wide attention [26–29]. However, already an ill-posed problem — since both the appearance models of parts and their instances in given images are unknown — it is exacerbated by the large viewpoint variation inherent in scenes. Instead, we employ a generic hand-crafted texture projection model to perform appearance and projective invariant detection of a wide range of meaningful textured scene regions.

Finally, our work is fundamentally different from that on **invariant texture description or recognition** based on hand-crafted descriptors [30] or by training classifiers for semantic or material properties of texture [31]. Where that line of work is focused on recognizing a wide range of *generic* texture from *cropped* images, we aim to *detect* a *specified* form of texture in indoor scenes, identify and address the challenges therein, and to explore how it helps high-level scene understanding tasks. We also differ from work that aims to learn to predict the presence or absence of generic material attributes in scenes [32].

3 Main Framework

3.1 Texture Frequency Projection Model

Shape-from-texture relates texture surface coordinates to corresponding camera coordinates in terms of the slant and tilt of the tangent plane at a point [16,33], or in terms of the plane gradients or normal [19,21,34]. Surface coordinates (expressed in camera reference frame) are then projected to the image plane via scaled orthographic or perspective projection, assuming the camera focal length is known. Since we are interested in planar rectification, we can relate the surface and image points via a planar homography. This does not require the focal length, but the downside, as we shall see shortly, is that the rectification is only up to an affine ambiguity. Let us represent the projective transform from the image plane to the textured surface plane as a 3×3 homography H, i.e., $\mathbf{x}'_\mathbf{s} = H\mathbf{x}_\mathbf{i}$ (see Fig. 3). H can be decomposed to separate the contributions of the affine part and the projective part [35]:

$$H = H_A H_P = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ h_7 & h_8 & 1 \end{pmatrix} \tag{1}$$

That is, the image coordinates $\mathbf{x}_\mathbf{i}$ are first transformed by the "purely" projective homography (i.e. what is left in the projective group after removing the affine group) to some intermediate plane coordinates $\mathbf{x}_\mathbf{s} = (x_s \; y_s)^T$, followed by the affine transform H_A to obtain the world (fronto-parallel) plane coordinates

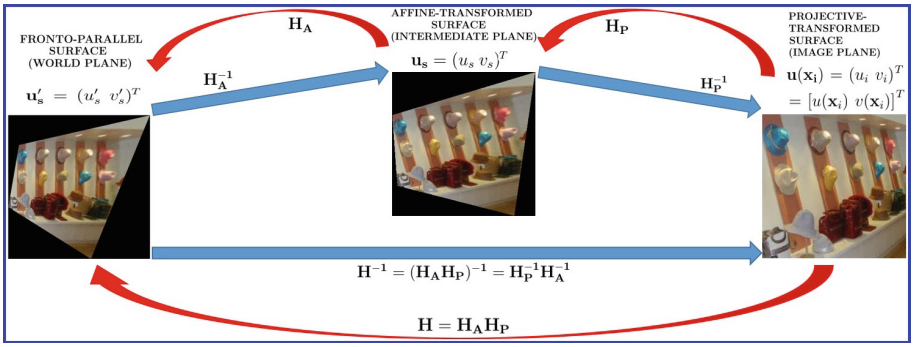

Fig. 3. Assorted hats along the bottom clutter this MIT Indoor67 `clothingstore` image (right), yet the *texture* is correctly affine-rectified by the proposed approach (center). For illustration, metric rectification (left) was manually performed, removing any rotation or anisotropic scaling.

$\mathbf{x}'_\mathbf{s} = (x'_s\ y'_s)^T$. Note that the last row of H_P is the same as the last row of H. We consider the role of H_A first, which provides the transformation:

$$x'_s = a_{11}x_s + a_{12}y_s + a_{13}, \qquad y'_s = a_{21}x_s + a_{22}y_s + a_{23} \tag{2}$$

The transpose of the Jacobian of H_A, given as:

$$\mathbf{J_A^T} = \begin{pmatrix} \frac{\partial x'_s}{\partial x_s} & \frac{\partial y'_s}{\partial x_s} \\ \\ \frac{\partial x'_s}{\partial y_s} & \frac{\partial y'_s}{\partial y_s} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} \\ \\ a_{12} & a_{22} \end{pmatrix} \tag{3}$$

transforms a world plane spatial frequency $\mathbf{u}'_\mathbf{s} = (u'_s\ v'_s)^T$ — which is constant over the entire plane, since we have assumed homogeneity of texture on the surface — into the frequency $\mathbf{u_s} = (u_s\ v_s)^T = \mathbf{J_A^T u}'_\mathbf{s}$ on our intermediate plane (c.f., [16]). Clearly, frequency $\mathbf{u_s}$ on the intermediate plane, albeit different from world plane frequency $\mathbf{u}'_\mathbf{s}$, is also constant, i.e., does not vary spatially. In other words, homogeneous texture upon affine transform remains homogeneous, as also observed in [22]. A similar analysis for H_P, which transforms image points $\mathbf{x_i} = (x_i\ y_i)^T$ into points $\mathbf{x_s} = (x_s\ y_s)^T$ on our intermediate plane, gives:

$$\mathbf{J_P^T} = \begin{pmatrix} \frac{\partial x_s}{\partial x_i} & \frac{\partial y_s}{\partial x_i} \\ \\ \frac{\partial x_s}{\partial y_i} & \frac{\partial y_s}{\partial y_i} \end{pmatrix} = \frac{1}{(h_7 x_i + h_8 y_i + 1)^2} \begin{pmatrix} h_8 y_i + 1 & -h_7 y_i \\ -h_8 x_i & h_7 x_i + 1 \end{pmatrix} \tag{4}$$

$\mathbf{J_P^T}$ transforms the intermediate plane constant frequency $\mathbf{u_s} = (u_s\ v_s)^T$ to image plane variable frequency $\mathbf{u}(\mathbf{x_i}) = (u_i\ v_i)^T = [u(\mathbf{x}_i)\ v(\mathbf{x}_i)]^T = \mathbf{J_P^T u_s}$. While the above analysis is applicable to *any* spatial frequency component, in Sect. 3.2 we shall obtain a robust instantaneous estimate of the *dominant* spatial frequency component in a given image patch depicting real-world texture,

which inevitably contains multiple frequency components. Denote this estimate as $\tilde{\mathbf{u}}(\mathbf{x_i}) = (\tilde{u}_i\ \tilde{v}_i)' = [\tilde{u}(\mathbf{x}_i)\ \tilde{v}(\mathbf{x}_i)]'$. We then arrive at a method to recover H_P by minimizing the following ***re-projection error*** $E_{RP}(h_7, h_8, u_s, v_s)$ over the projective parameters h_7, h_8 and the intermediate plane frequency u_s, v_s:

$$E_{RP} = \sum_{x_i}\sum_{y_i}\left(\frac{(h_8 y_i + 1)u_s - h_7 y_i v_s}{(h_7 x_i + h_8 y_i + 1)^2} - \tilde{u_i}\right)^2 + \left(\frac{(h_7 x_i + 1)v_s - h_8 x_i u_s}{(h_7 x_i + h_8 y_i + 1)^2} - \tilde{v_i}\right)^2 \quad (5)$$

Optimizing Eq. 5 is a nonlinear least squares problem, and we solve it using the Levenberg-Marquardt algorithm. Observe that our method allows the recovery of H_P and not H_A. This is because $\mathbf{J_A^T}$ maps the fronto-parallel plane frequency $\mathbf{u'_s} = (u'_s\ v'_s)^T$ to a different but still constant frequency $\mathbf{u_s} = (u_s\ v_s)^T$. As such, a planar rectification only to within an ambiguous affine transform H_A^{-1} of the fronto-parallel plane may be obtained.

3.2 Optimal Estimation of Dominant Frequency in Projected Homogeneous Texture

A Gabor filter $h(\mathbf{u}; \mathbf{x})$ with center frequency $\mathbf{u} = (u, v)$ can be convolved with an image $f(\mathbf{x})$ to give its frequency content near $\mathbf{u}$ at point $\mathbf{x} = (x, y)$:

$$A(\mathbf{u}; \mathbf{x}) = |f(\mathbf{x}) * h(\mathbf{u}; \mathbf{x})| \quad (6)$$

Since a given texture may exhibit multiple frequencies, which may also be oriented differently, one must discern the component that can be reliably tracked over the image, so as to be able to use the projection model developed in Sect. 3.1. In this regard, Super and Bovik [16] have previously demonstrated estimation of the *dominant* texture frequency — a distinct peak at any given point, around which most of the energy is concentrated in a narrow band — employing a frequency demodulation model (**DEMOD**) from [20]. Briefly, denote the horizontal and vertical partial derivatives of Gabor filter $h(\mathbf{u}; \mathbf{x})$ by $h_x(\mathbf{u}; \mathbf{x})$ and $h_y(\mathbf{u}; \mathbf{x})$, respectively, and the corresponding amplitude response (Eq. 6) by $B(\mathbf{u}; \mathbf{x})$ and $C(\mathbf{u}; \mathbf{x})$, respectively. Then, an *unsigned* instantaneous estimate $|\tilde{\mathbf{u}}(\mathbf{x})|$ of the *dominant* frequency component may be computed for the filter h that maximizes the response $A(\mathbf{u}; \mathbf{x})$ at each point as:

$$|\tilde{u}(\mathbf{x})| = \frac{B(\mathbf{u}; \mathbf{x})}{2\pi A(\mathbf{u}; \mathbf{x})}, \qquad |\tilde{v}(\mathbf{x})| = \frac{C(\mathbf{u}; \mathbf{x})}{2\pi A(\mathbf{u}; \mathbf{x})} \quad (7)$$

The sign at each pixel is defined by the frequency plane quadrant wherefrom the maximizing Gabor is sampled. Only quadrants I or IV are used, since the Fourier spectrum is symmetric.

Frequency Drift. For the MIT Indoor67 `airport_inside` image patch shown in Fig. 4(a), DEMOD [16] provides a rather poor estimate of the dominant frequency, resulting in poor rectification using the model from Sect. 3.1. This is not

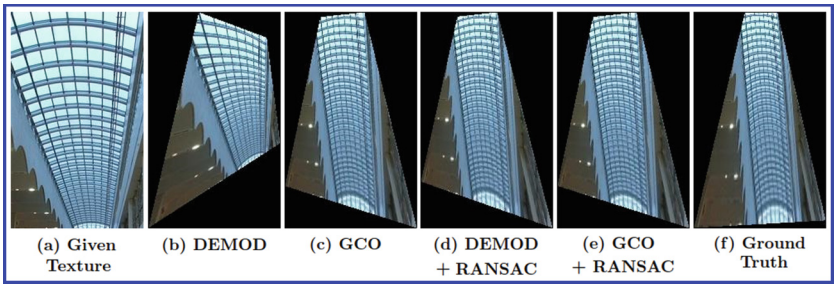

Fig. 4. Affine rectification of texture (a) via the model developed in Sect. 3.1, applied to non-optimal frequency estimate by DEMOD [16] is prone to drift (b); optimal frequency estimation via GCO improves performance (c).

surprising given the grim challenges we outlined in Sect. 1. Figure 5 examines the dominant frequency estimates in detail. Since the given texture does not extend to the lower left and lower right regions in the image patch, the maximizing Gabor *drifts* in *both* the center frequency (Fig. 5(a)) and orientation (Fig. 5(c)) in these regions (brighter pixels depict numerically larger values). A 1D plot along the dotted line (Fig. 5(b)) shows the center frequency deviates in these regions from an otherwise increasing pattern. The orientation plot (Fig. 5(d)) reveals that the Gabors pre-dominantly fire strongly at the horizontal bars in the image (18°, 0°, −18° as one moves from left to right). However, in the lower region, it is the vertical bars (−72°, 90°) that define the "dominant" Gabors. Figures 5(e) and (f), respectively, show the resulting horizontal and vertical estimates obtained via Eqs. 7. Corresponding surface plots are depicted in Figs. 5(f) and (h), showing large discontinuities. We propose to resolve drift by enforcing smoothness via the following regularized graph cut problem [36]:

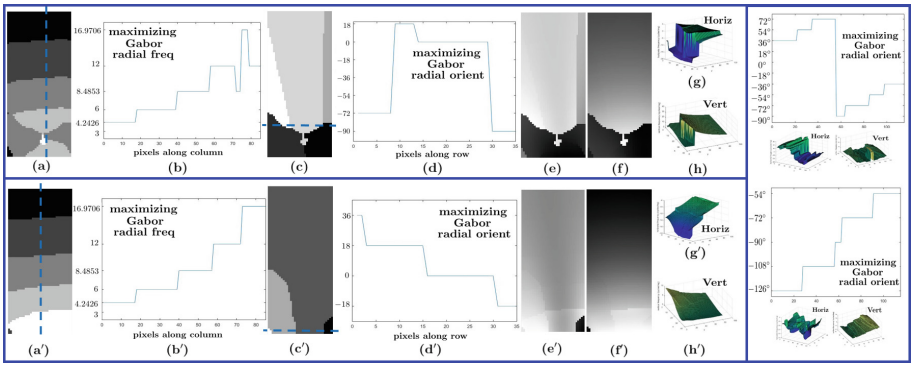

Fig. 5. TOP: Closer look at drift in dominant instantaneous frequency estimate via DEMOD [16]. **BOT:** GCO resolves drift in both center radial frequency and orientation. **Right:** GCO also resolves quadrant ambiguity, if any.

$$E(f) = \sum_{p \in \mathcal{P}} D_p(f_p) + \sum_{\{p,q\} \in \mathcal{N}} V_{p,q}(f_p, f_q) \tag{8}$$

where $\mathcal{P}$ is the set of sites p to be labeled (pixels), and $\mathcal{N}$ is the 8-neighbourhood system. Our set of labels $\mathcal{L}$ is the Gabor filter bank. The unary term is defined as $D_p(f_p) = \alpha / A(f_p; p)$, where $A(\mathbf{u}; \mathbf{x})$ is as dictated by Eq. 6, $\alpha = 1$ and $f_p = (\Omega_p, \theta_p) \in \mathcal{L}$ gives the filter with center frequency $\mathbf{u} = (\Omega_p sin\theta_p,\ \Omega_p cos\theta_p)$ at $\mathbf{x} = p$. The pairwise term $V_{p,q}(f_p, f_q) = V(f_p, f_q)$ forces the center radial frequency Ω_p and orientation θ_p to be smooth:

$$V(f_p, f_q) = \beta(\Omega_p - \Omega_q)^2 + \gamma\{(sin\theta_p - sin\theta_q)^2 + (cos\theta_p - cos\theta_q)^2\} \tag{9}$$

Demodulation (Eqs. 7) is then performed *after* solving Eq. 8 for the optimal labeling f. We call this scheme Graph Cut Optimization (GCO), solved via α-expansion [36]. As seen in Fig. 5(bottom) it yields a smooth, monotonically increasing frequency and orientation profile, consequently providing an improved rectification (Fig. 4(c)) compared to the non-optimal case (Fig. 4(b)).

A workaround to drift is to perform a robust parameter estimation via RANSAC [37]. While this can seemingly handle drift (see Fig. 4(d)), the %outliers is significantly higher compared to when GCO is also used in conjunction with RANSAC (Table 1). Later in Sect. 4.2, we employ the %outliers as a metric to "detect" homogeneous texture, and since GCO renders %outliers a more adequate measure, it is indispensable if we wish to reliably differentiate between non-textured surfaces from textured surfaces perturbed by other scene elements.

Table 1. Recovered projective parameters and % outliers for the example texture in Fig. 4(a) via DEMOD, GCO, DEMOD+RANSAC and GCO+RANSAC

	DEMOD	GCO	DEMOD+RANSAC	GCO+RANSAC	GT
h7	0.2940	−0.0736	−0.0750	−0.0733	0.0089
h8	−0.2650	−0.4923	−0.5267	−0.4962	−0.6035
% outliers	N/A	N/A	10.36 %	3.63 %	N/A

Quadrant Ambiguity. DEMOD [16] also fails on, e.g., the `subway` patch in Fig. 6(o), because it can only measure the frequency orientations modulo-π (frequency estimates from opposite quadrants have the *same* magnitude). This wrapped orientation may result in sharp discontinuity between neighboring frequency estimates. As explained in Fig. 5(top right), the orientation of the rails increases as one moves from left to right (36°, 54°, 72°), and wraps around back to −90°. We extend our set of labels $\mathcal{L}$ to include filters sampled at orientations from *all* the four quadrants, and rely on the smoothness constraint between neighboring pixels to resolve the quadrant ambiguity. As seen in Fig. 5(lower right), the optimal orientations recovered by GCO are those sampled from quadrant III and not I, thereby ensuring a smoother transition into quadrant IV with respect to both the demodulated horizontal and vertical frequency.

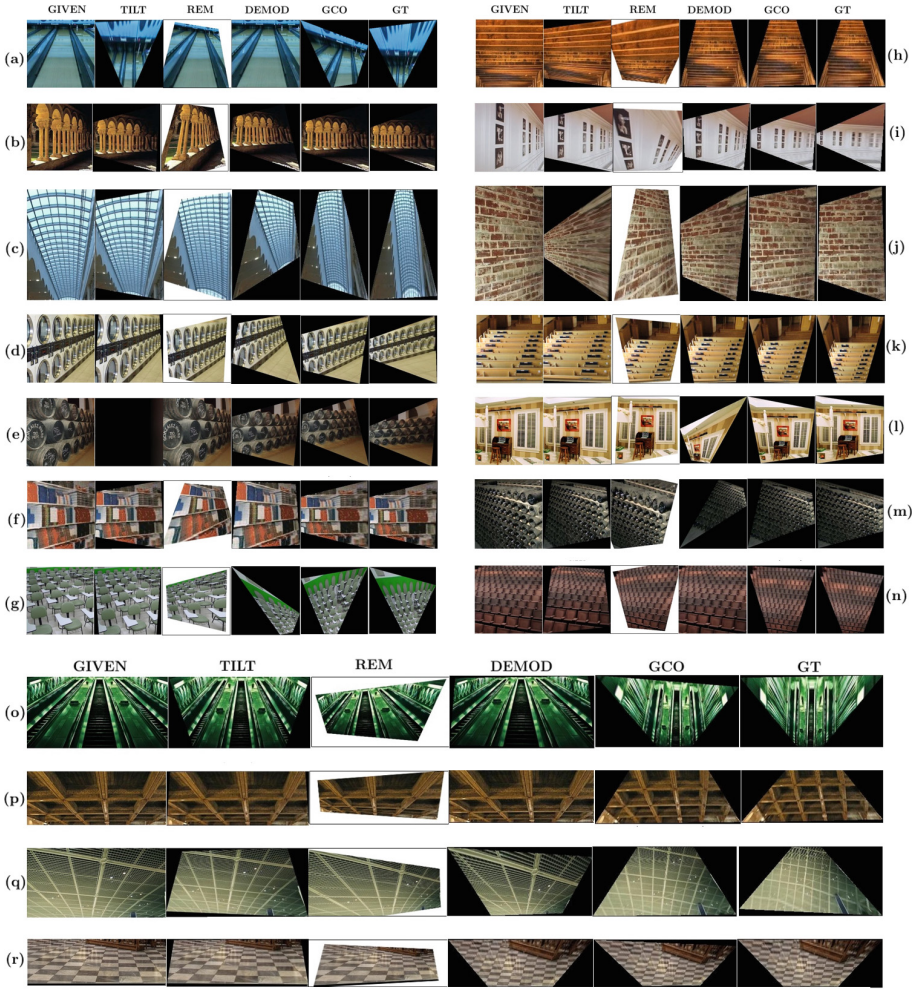

Fig. 6. Affine rectification of homogeneous texture: Given, TILT [15], REM [6], DEMOD [16] with our model in Sect. 3.1, Proposed (GCO) and Ground Truth.

4 Experiments

4.1 Affine Rectification

The proposed affine rectification is evaluated on $N = 30$ patches, cropped from various images in MIT Indoor67, depicting some homogeneous texture under perspective projection. We compare with TILT (Transform-Invariant Low-rank Texture) [15] using publicly available code, with REM (Repetition Maximization) [6] using their command-line tool, and our implementation of DEMOD [16] in conjunction with our model from Sect. 3.1, thereby encompassing techniques based on low-rankness, recurring elements and frequency. Following TILT

and REM, a multi-scale approach is also implemented for the proposed GCO scheme (see supple. material). We define the **mean estimation error** as $\sum_{i=1}^{N} \sqrt{(\tilde{h_{7i}} - h_{7i})^2 + (\tilde{h_{8i}} - h_{8i})^2}$, where $\tilde{h_{7i}}$, $\tilde{h_{8i}}$ are the parameters returned by an algorithm, and h_7, h_8 are the ground truth parameters obtained by manual annotation of vanishing points. The various algorithms fare as follows: TILT: 0.496, DEMOD: 0.386, and GCO: **0.186**. REM does not return the estimated parameters, hence its performance is not quantified. Our proposed GCO has substantially improved upon the pure DEMOD. TILT performs even worse than DEMOD, but of that, more later.

Figure 6 presents some qualitative results. REM — which has only been demonstrated for properly cropped, printed patterns in [6] — seems to only perform in the infrequent cases where it can detect some regular lattice structure (e.g., k, l), but usually either produces a partial rectification (c, n), or fails altogether. TILT, in general, also performs well only on a few cases, where the underlying texture is low-rank (a, b), but breaks down when this assumption is violated — e.g., port-holes (d), or barrels (e), where the gradients are isotropic in all directions. On the other hand, our robustified frequency based scheme (GCO) is seen to handle such texture very well, corroborating our intuition that homogeneity is a more general assumption than low-rankness. TILT and REM also seem to fail on cases exhibiting large perspective distortion — e.g., the textured ceilings in cases (p, q) — and when illumination changes over the texture (m, o, r). On the other hand, use of Gabor filters allows our frequency based scheme to perform remarkably well in these challenging cases. Provided the scale of texture is small (i.e., texture contains higher frequencies) relative to the scale of the surface it covers, a frequency based representation is resilient to slow-varying (low-frequency) photometric changes [16].

4.2 Detection in the Wild

Overlapping patches (80×80 pixels) are sampled over a multi-scale image pyramid (details in supple. material) to decide if they are textured planar patches or not. GCO and robust parameter estimation via RANSAC (with outlier threshold fixed at 0.01) is performed on each patch individually. Our error measure (Eq. 5) is not affine invariant, so we employ the following heuristic normalization. First, the dynamic range of the optimal radial frequency ($\tilde{\mathbf{u}}_\mathbf{i} = \tilde{\mathbf{u}}(\mathbf{x_i}) = \sqrt{\tilde{u_i}^2 + \tilde{v_i}^2}$) of RANSAC inliers is computed as $\mathcal{DR} = \max_{i \in inliers}(\tilde{\mathbf{u}}_\mathbf{i}) - \min_{i \in inliers}(\tilde{\mathbf{u}}_\mathbf{i})$. A normalized residual error is then computed for all pixels (i.e., inliers and outliers) as $\mathcal{E}(\mathbf{x_i}) = \{\tilde{\mathbf{u}}(\mathbf{x_i}) - \mathbf{J_P^T}(\mathbf{x_i})\mathbf{u_s}\}/\mathcal{DR}$, followed by re-evaluating %outliers (which serves as the decision score).

A quantitative evaluation is performed on 300 images sampled from the MIT Indoor67 (with at least 3 from each scene category) that have been manually annotated with quadrilaterals indicating the homogeneous textured regions, their plane projective parameters, and their semantic class IDs (left/right wall, ceiling, floor). We define true positives (TP), false positives (FP) and false negatives

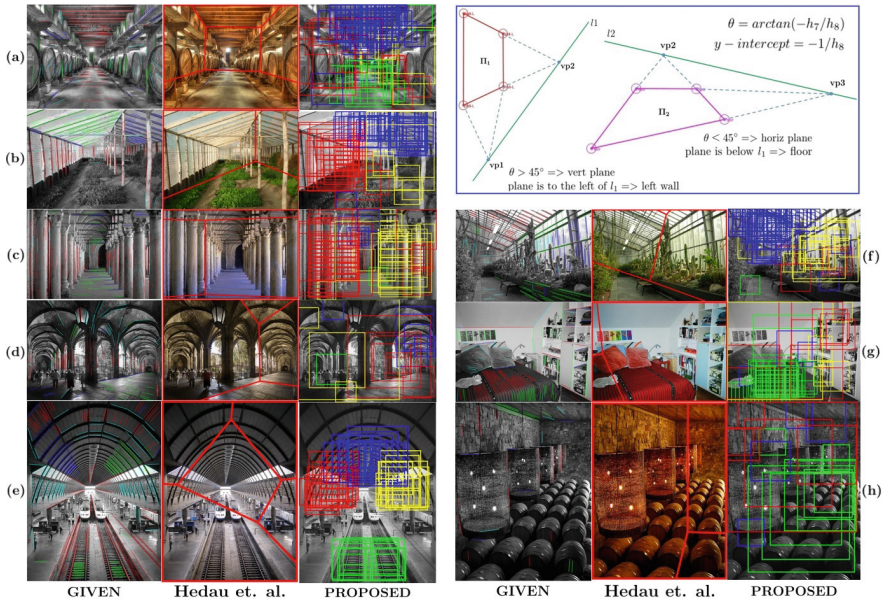

Fig. 7. Scene layout estimation by homogeneous texture detections, and associated vanishing lines. Scene with vanishing point clusters (left), box layouts [5] (center), proposed (right). Left wall = red, right wall = yellow, ceiling = blue, floor = green. **Best viewed in color.** (Color figure online)

(FN) as follows.[1] For **precision** [TP/(TP+FP)], TP is the number of candidate patches whose estimated semantic class (see Sect. 4.3) matches with an annotated region, with 50 % intersection-over-detection (i.e., at least 50 % of the candidate's area should cover the annotation), while FP is the number of candidates that fail this criterion. For **recall** [TP/(TP+FN)], TP is the number of annotated regions that are "fired on" by one or more candidates (with the correct semantic class), such that its area beyond a certain threshold is covered (we evaluated at both coverage >= 50 % and >= 80 %), while FN is the number of our annotated regions that fail this criterion. Note that for recall, TP + FN = 1367, which is the total number of annotated regions, similar to object detection [38].

Figure 8 presents precision-recall curves, and recall vs. # proposals curves for our method, as well as for TILT [15] (using ratio of final to initial rank as a decision score). One can observe a considerably superior performance by our method, with an **average precision = 0.53**, compared to 0.15 by TILT. Both methods improve in recall with increasing #proposals, but ours is seen to maintain a larger recall for the same #proposals from the outset. Some qualitative results are presented in Fig. 2 (and many more in supple. material).

[1] Since our detector is not "trained" to produce an exact bounding box, we somewhat differ in our definitions of these parameters from object detection [38]. Object detection methodology considers any more than one detection for a given ground truth as FPs, but all such detections are considered TPs in our scenario.

4.3 Indoor Scene Geometric Layout Estimation

Hedau et al. [5] have previously demonstrated the estimation of indoor scene geometric layout by using orthogonal vanishing points [39] to establish room orientation, and then using machine learning with rich feature sets [40] to localize room faces (i.e., ceiling, walls, floor) in space and scale. Figure 7 identifies the shortcomings of such an approach, using MIT Indoor67 images. Presence of more than three dominant planar directions (b, f, g), absence of straight lines in a certain direction (c), forked layout (d), and non-Manhattan structure (commonplace in real-world scenes) (e) are scenarios where such a scheme is apt to provide incorrect room orientation, while face localization is also prone to error (a, h) owing to the limitations of a learning based system, such as non-exhaustive training data.

Our detections and the recovered projective parameters provide an alternative scheme to estimate indoor geometric layout in textured scenes (Fig. 7), that requires neither vanishing points nor machine learning. A given detection may be classified as a vertical/horizontal surface depending on the slope of the vanishing line, and as left/right wall or ceiling/floor depending on the position of this line with respect to the patch center (see top right of Fig. 7 for details). The top 150 detections (sorted by % of RANSAC outliers) are then subjected to non-max suppression (NMS) performed *across* semantic classes (i.e., an incoming detection is not admitted if at least 50 % of its area is already occupied by *any* previously admitted and thus higher-ranked patch that is *not* from the same class). Of course, the proposed scheme requires the scene faces to be textured. For e.g., Fig. 7(g) shows a scenario where the non-textured ceiling or walls cannot be correctly assigned a semantic face category.

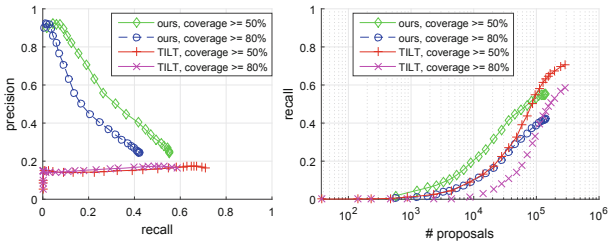

Fig. 8. Precision-recall and recall vs. # proposals.

Fig. 9. Sample correct classifications.

4.4 Indoor Scene Classification

Table 2 quantitatively demonstrates that affine rectification of textured patches detected (with decision threshold fixed at 50 % RANSAC outliers) via the proposed approach can improve scene classification performance. Best practices for dense local feature based classification as suggested in [41] are followed (details in supple. material), using Fisher Encoding with sum pooling [42], Hellinger

Table 2. L: MIT Indoor67 classification improvement with fisher encoding of dense descriptors (CENTRIST [44], LBP [45], SIFT [46,47], HOG2 × 2 [48,49]) extracted from affine-rectified texture patches. **R**: State of the art performance — *all* (except SIFT [28]) involve learning-based feature extraction, unlike ours

Representation (Ours)	**% Accuracy**
LBP_u2(16,2)	37.10%
LBP_u2_Rect(16,2)	**40.84%**
LBP_u2 + LBP_u2_Rect	**41.28%**
CEN	46.44%
CEN_Rect	46.30%
CEN + CEN_Rect	**50.22%**
SIFT	59.14%
SIFT_Rect	57.98%
SIFT + SIFT_Rect	**61.00%**
HOG	57.69%
HOG_Rect	56.65%
HOG + HOG_Rect	**60.42%**
CEN + SIFT + HOG	61.66%
SIFT_Rect + HOG_Rect	60.88%
CEN + SIFT + HOG + SIFT_Rect + HOG_Rect	**64.54%**

Single Rep.(State of Art)	**% Accuracy**
OPM [50]	51.45%
Mode Seeking [29]	**64.03%**
SIFT [28]	60.77%
BoP [28]	46.10%
DSFL [51]	52.24%
DeCAF [51] (deep learn.)	58.52%
MOP-CNN [52] (deep learn.)	**68.88%**
Combined Rep.(State of Art)	**% Accuracy**
BoP + SIFT [28]	63.10%
OMP + SPM [50]	63.48%
Mode Seeking + SIFT [29]	66.87%
ISPR + SIFT [53]	**68.5%**
SIFT + DeCAF [51] (deep learn.)	70.51%
DSFL + DeCAF [51] (deep learn.)	**76.23%**

Kernel mapping, one-vs-all linear SVMs [43], and various gradient and thresholding based descriptors. Both regular, as well as rectified representations (wherein dense descriptors are extracted from affine-rectified patches) are computed, and then combined via the score fusion scheme of [26].

In general, our rectification based representations, on their own, perform slightly lower than regular ones since descriptors are extracted only from *detected* textured regions, which, more often than not, span the image only in some spatial regions and at certain scales, and not exhaustively, therein losing some discriminative power. Interestingly, however, LBP, perhaps because it is inherently a *texture* descriptor, performs significantly better with rectified, detected texture. For similar reasons, both representations perform almost the same with CENTRIST — again, a texture descriptor. Finally, our results suggest that features extracted upon planar rectification are also complementary to regular features, a finding that is consistent across all the descriptors experimented with. Figure 9 shows some sample images that were mis-classified using a regular representation (SIFT+HOG), but were correctly classified using (SIFT_Rect +HOG_Rect). A notable property among most of them is the presence of large perspective distortion, as well as high-frequency homogeneous texture.

5 Conclusion

This paper has demonstrated a projective-invariant method to detect homogeneous texture, as well as to perform its affine rectification in challenging, real-world indoor scenes, outperforming existing representative work. Homogeneous texture is seen to provide cues for indoor geometric layout estimation in scenes

where vanishing points cannot be reliably computed or the Manhattan assumption is violated. Rectified homogeneous texture also facilitates improved indoor scene recognition on the MIT Indoor67 benchmark, demonstrating that plane projective rectification can push performance in a recognition system.

References

1. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: CVPR (2009)
2. Picard, R.W.: A society of models for video and image libraries. IBM Syst. J. **35**(3.4), 292–312 (2010)
3. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. IJCV **65**(1–2), 43–72 (2005)
4. Coughlan, J.M., Yuille, A.L.: Manhattan world: compass direction from a single image by Bayesian inference. In: ICCV (1999)
5. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: ICCV (2009)
6. Aiger, D., Cohen-Or, D., Mitra, N.J.: Repetition maximization based texture rectification. EUROGRAPHICS **31**(2pt2), 439–448 (2012)
7. Chum, O., Matas, J.: Planar affine rectification from change of scale. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part IV. LNCS, vol. 6495, pp. 347–360. Springer, Heidelberg (2011)
8. Leung, T., Malik, J.: Detecting, localizing and grouping repeated scene elements from an image. In: Buxton, B., Cipolla, R. (eds.) Computer Vision — ECCV 1996. LNCS, vol. 1064, pp. 546–555. Springer, Heidelberg (1996)
9. Pritts, J., Chum, O., Matas, J.: Detection, rectification and segmentation of coplanar repeated patterns. In: CVPR (2014)
10. Schaffalitzky, F., Zisserman, A.: Geometric grouping of repeated elements withinimages. In: BMVC (1998)
11. Wu, C., Frahm, J.-M., Pollefeys, M.: Detecting large repetitive structures with salient boundaries. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 142–155. Springer, Heidelberg (2010)
12. Wu, C., Frahm, J.M., Pollefeys, M.: Repetition-based dense single-view reconstruction. In: CVPR (2011)
13. Hong, W., Yang, A.Y., Huang, K., Ma, Y.: On symmetry and multiple-view geometry: Structure, pose, and calibration from a single image. IJCV **60**(3), 241–265 (2004)
14. Tuytelaars, T., Turina, A., Gool, L.V.: Noncombinatorial detection of regular repetitions under perspective skew. TPAMI **25**(4), 418–432 (2003)
15. Zhang, Z., Liang, X., Ganesh, A., Ma, Y.: TILT: transform invariant low-rank textures. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part III. LNCS, vol. 6494, pp. 314–328. Springer, Heidelberg (2010)
16. Super, B.J., Bovik, A.C.: Planar surface orientation from texture spatial frequencies. Pattern Recogn. **28**(5), 729–743 (1995)
17. Rosenholtz, R., Malik, J.: Surface orientation from texture: isotropy or homogeneity (or both)? Vis. Res. **37**(16), 2283–2293 (1997)
18. Ribeiro, E., Hancock, E.R.: Estimating the 3D orientation of texture planes using local spectral analysis. Image Vis. Comput. **18**(8), 619–631 (2000)

19. Super, B.J., Bovik, A.C.: Three-dimensional orientation from texture using gabor wavelets. In: Proceedings of the SPIE Visual Communications and Image Processing 1991: Image Processing (1991)
20. Havlicek, J.P., Bovik, A.C., Maragos, P.: Modulation models for image processing and wavelet-based image demodulation. In: Proceedings of the Asilomar Conference on Signals, Systems and Computers (1992)
21. Collins, T., Durou, J., Gurdjos, P., Bartoli, A.: Single-view perspective shape-from-texture with focal length estimation: a piecewise affine approach. In: Proceedings of the 3D Data Processing, Visualization and Transmission (3DPVT) (2010)
22. Criminsi, A., Zisserman, A.: Shape from texture: homogeneity revisited. In: BMVC (2000)
23. Shaw, D., Barnes, N.: Perspective rectangle detection. In: European Conference on Computer Vision Workshop on Applications of Computer Vision (2006)
24. Kosecka, J., Zhang, W.: Extraction, matching and pose recovery based on dominant rectangular structures. In: First IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis, 2003 (2003)
25. Stella, X.Y., Zhang, H., Malik, J.: Inferring spatial layout from a single image via depth-ordered grouping. In: CVPR Workshop (2008)
26. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: ICCV (2011)
27. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 73–86. Springer, Heidelberg (2012)
28. Juneja, M., Vedaldi, A., Jawahar, C.V., Zisserman, A.: Blocks that shout: distinctive parts for scene classification. In: CVPR (2013)
29. Doersch, C., Gupta, A., Efros, A.A.: Mid-level visual element discovery as discriminative mode seeking. In: Proceedings of the Neural Information Processing Systems (2013)
30. Zhang, J., Marszaek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. IJCV **73**(2), 213–238 (2007)
31. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: CVPR (2014)
32. Patterson, G., Xu, C., Su, H., Hays, J.: The SUN attribute database: beyond categories for deeper scene understanding. IJCV **108**(1), 59–81 (2014)
33. Super, B.J., Bovik, A.C.: Shape from texture using local spectral moments. TPAMI **17**(4), 333–343 (1995)
34. Krumm, J., Shafer, S.: Shape from periodic texture using the spectrogram. In: CVPR (1992)
35. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004). ISBN 0521540518
36. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. TPAMI **23**(11), 1222–1239 (2001)
37. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
38. Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes challenge: a retrospective. IJCV **111**(1), 98–136 (2014)

39. Rother, C.: A new approach for vanishing point detection in architectural environments. In: BMVC (2000)
40. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. IJCV **75**(1), 151–172 (2007)
41. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC (2011)
42. Vedaldi, A., Fulkerson, B.: VLFeat: an open and portable library of computer vision algorithms (2008). http://www.vlfeat.org/
43. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM TILT **2**, 27:1–27:27 (2011). http://www.csie.ntu.edu.tw/cjlin/libsvm
44. Wu, J., Rehg, J.M.: CENTRIST: a visual descriptor for scene categorization. TPAMI **33**(8), 1489–1501 (2011)
45. Ojala, T., Pietikinen, M., Menp, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. TPAMI **24**(7), 971–987 (2002)
46. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2), 91–110 (2004)
47. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: CVPR (2006)
48. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: ICCV (2005)
49. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI **32**(9), 1627–1645 (2010)
50. Xie, L., Wang, J., Guo, B., Zhang, B., Tian, Q.: Orientational pyramid matching for recognizing indoor scenes. In: CVPR (2014)
51. Zuo, Z., Wang, G., Shuai, B., Zhao, L., Yang, Q., Jiang, X.: Learning discriminative and shareable features for scene classification. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part I. LNCS, vol. 8689, pp. 552–568. Springer, Heidelberg (2014)
52. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VII. LNCS, vol. 8695, pp. 392–407. Springer, Heidelberg (2014)
53. Lin, D., Lu, C., Liao, R., Jia, J.: Learning important spatial pooling regions for scene classification. In: CVPR (2014)