Detecting Vanishing Points in Natural Scenes with Application in Photo Composition Analysis

Zihan Zhou, Farshid Farhat, and James Z. Wang

***Abstract*—Linear perspective is widely used in landscape photography to create the impression of depth on a 2D photo. Automated understanding of the use of linear perspective in landscape photography has a number of real-world applications, including aesthetics assessment, image retrieval, and on-site feedback for photo composition. We address this problem by detecting vanishing points and the associated line structures in photos. However, natural landscape scenes pose great technical challenges because there are often inadequate number of strong edges converging to the vanishing points. To overcome this difficulty, we propose a novel vanishing point detection method that exploits global structures in the scene via contour detection. We show that our method significantly outperforms state-of-the-art methods on a public ground truth landscape image dataset that we have created. Based on the detection results, we further demonstrate how our approach to linear perspective understanding can be used to provide on-site guidance to amateur photographers on their work through a novel viewpoint-specific image retrieval system.**

***Index Terms*—Vanishing Point; Photo Composition; Image Retrieval.**

I. Introduction

RECENTLY, with the widespread use of digital cameras and other mobile imaging devices, there has been increasing interest in the multimedia community in building intelligent programs to automatically analyze the visual aesthetics and composition of photos. Information about photo aesthetics and composition [1] is shown to benefit many real-world applications. For example, it can be used to suggest improvements to the aesthetics and composition of photographers' work through image re-targeting [2], [3], as well as provide on-site feedback to the photographer at the point of photographic creation [4], [5].

In this paper, we focus on an important principle in photo composition, namely, the use of *linear perspective*. It corresponds to a relative complex spatial system that concerns primarily with the *parallel lines* in the scene. Indeed, parallel lines are one of the most prevalent geometric structures in both man-made and natural environments. Under the pinhole camera model, they are projected into image lines which converge to a single point, namely, the vanishing point (VP). Because the VPs provide crucial information about the geometric structure of the scene, automatic detection of VPs have long been an active research problem in image understanding.

Z. Zhou and J. Z. Wang are with College of Information Sciences and Technology, The Pennsylvania State University, USA (e-mail: zzhou@ist.psu.edu; jwang@ist.psu.edu.). F. Farhat is with the Department of Computer Science and Engineering, The Pennsylvania State University, USA (e-mail: fuf111@cse.psu.edu).

Fig. 1. Natural scene images with vanishing points. Images are from the "landscape" category of the AVA dataset. Manually labeled ground truth lines are marked in green.

In the literature, existing VP detection methods mainly focus on the *man-made environments*, which typically consist of a large number of edges or line segments aligned to one or more dominant directions. To this end, numerous methods have been proposed to cluster line segments into groups, each representing a VP in the scene [6], [7], [8], [9], [10], [11]. These methods have successfully found real-world applications such as self-calibration, 3D reconstruction of urban scenes, and stereo matching.

However, little attention has been paid to the *natural landscape scenes* – a significant genre in both professional and consumer photography. In natural scene images, a VP is detectable when there are as few as two parallel lines in space. As shown in Figure 1, such VPs and the associated geometric structures convey a strong sense of 3D space or depth to the viewers. While human eyes have little difficulty identifying the VPs in these images, automatic detection of VPs poses great challenge to computer systems for two main reasons. *First, the visible edges can be weak and not detectable via local photometric cues.* Existing line segment detection methods typically assume the gradient magnitude of an edge pixel is above a certain threshold (*e.g.*, Canny edge detector [12]) or the number of pixels with aligned gradient orientations is above a certain threshold (*e.g.*, LSD [13]). However, determining the threshold can be difficult due to the weak edges and image noise. *Second, the number of edges converging to the VP may be small compared to irrelevant edges in the same scene.* As a result, even if one can detect the converging edges, clustering them into one group can be demanding due to the large number of outliers.

To overcome these problems, we propose to make use of the *image contours* to detect edges in natural images. Compared to local edge detectors, image contours encode valuable global information about the scene, thus are more effective in recognizing weak edges while reducing the number of false detections due to textures. By combining the contour-based edge detection with J-Linkage [14], a popular multi-model detection algorithm, our method has been shown to significantly outperforms state-of-the-art methods on detecting the dominant VP in natural scene images.

As an application of our VP detection method, we demonstrate how the detected VPs can be used to improve the usefulness of existing content-based image retrieval systems in providing on-site feedback to amateur photographers. In particular, we note that linear perspective is known as an effective tool in *recognizing the viewer as a specific unique individual in a distinct place with a point of view* [15]. Therefore, given a photo taken by the user, we study the problem of finding photos about similar scenes and with *similar viewpoints* in a large collection of photos. These photos can be potentially used to provide guidance to the user on his own work. Further, in this task, we are also the first to answer an important yet largely unexplored question in the literature: *How to determine whether there exists a dominant VP in a photo?* To this end, we design a new measure of strength for a given candidate VP, and systematically examine its effectiveness on our dataset.

In summary, the main contributions are as follows.

- We propose the research problem of detecting vanishing points in natural landscape scenes and a new method for dominant VP detection. By combining a contour-based edge detector with J-Linkage, our method significantly outperforms state-of-the-art methods for natural scenes.
- We develop a new strength measure for VPs and demonstrate its effectiveness in identifying images with a dominant VP.
- We demonstrate the application of our method for assisting amateur photographers at the point of photographic creation via viewpoint-specific image retrieval.
- To facilitate future research, we have created and made available a manually labeled dataset for dominant VPs in over 1,300 real-world natural scene images.

II. Related Work

A. Vanishing Point Detection

Most existing VP detection algorithms are based on clustering edges in the image according to their orientations [6], [7], [8]. In [10], Xu *et al.* studied various consistency measures between the VPs and line segments and developed a new method that minimizes the uncertainty of the estimated VPs. Lezama *et al.* [11] proposed to find the VPs via point alignments based on the *a contrario* methodology. Instead of using edges, Vedaldi and Zisserman proposed to detect VPs by aligning self-similar structures [16]. As we discussed before, these methods are designed for man-made environments. Identifying and clustering edges for VP detection in natural scene images still remain a challenging problem.

Recently, there is an increasing interest in exploiting special scene structures for VP detection. For example, many methods assume the "Manhattan World" model, indicating that three orthogonal parallel-line clusters are present [17], [18], [9], [19]. When the assumption holds, it is shown to improve the VP detection results. But such assumption is invalid for typical natural scenes. Other related work detect VPs in specific scenes such as unstructured roads [20], [21], [22], but it remains unclear how these methods can be extended to general natural scenes.

B. Photo Composition Modeling

Photo composition, which describes the placement or arrangement of visual elements or objects within the frame, has long been a subject of study in computational photography. A line of work concern themselves with known photographic rules and design principles, such as simplicity, depth of field, golden ratio, rule of thirds, and visual balance. Based on these rules, various image retargeting and recomposition tools have been proposed to improve the image quality [2], [3], [23], [24]. We refer readers to [25] for a comprehensive survey on this topic. However, the use of linear perspective has largely been ignored in the existing work. Compared to the aforementioned rules which mainly focus on the 2D rendering of visual elements, linear perspective enables photographers to convey the sense of 3D space to the viewers.

Recently, data-driven approaches to composition modeling have gained increasing attention in the multimedia community. These methods make use of community contributed photos to automatically learn composition models from the data. For example, Yan *et al.* [26] propose a set of composition features and learn a model for automatic removal of distracting content and enhancement of the overall composition. In [27], a unified photo enhancement framework is proposed based on the discovery and modeling of aesthetic communities on Flickr. Besides offline image enhancement, composition models learned from exemplar photos can also been used to provide online aesthetics guidance to the photographers, such as selecting the best view [28], [5], recommending the locations and poses of human subjects in a photograph [29], [30], [31], and suggesting the appropriate camera parameters (*e.g.*, aperture, ISO, and exposure) [32]. Our work also takes advantage of vast data available through photo sharing websites. But unlike existing work that each focuses on certain specific aspects of the photo, we take a completely different approach to on-site feedback and aim to provide comprehensive photographic guidance through a novel composition-sensitive image retrieval system.

C. Image Retrieval

The classic approaches to content-based image retrieval [33] typically measure the visual similarity based on low-level features (*e.g.*, color, texture, and shape). Recently, thanks to the availability of large-scale image datasets and computing resources, complicated models have been trained to capture the high-level semantics about the scene [34], [35], [36], [37]. However, because many visual descriptors are generated by

local feature extraction processes, the overall spatial composition of the image (*i.e.*, from which viewpoint the image is taken) is usually neglected. To remedy this issue, [4] first classify images into pre-defined composition categories such as "horizontal", "vertical", and "diagonal". Similar to our work, [38] also explore the VPs in the image for retrieval. However, it assumes known VP locations in all images, thus cannot be applied to general image database where majority of the images do not contain a VP.

III. Ground Truth Dataset

To create our ground truth dataset, we leverage the open AVA dataset [39], which contains over 250,000 images along with a variety of annotations. The dataset provides semantic tags describing the semantics of the images for over 60 categories, such as "natural", "landscape", "macro", and "urban". For this work, we used the 21,982 images labeled as "landscape".

Next, for each image, we need to determine whether it contains a dominant VP and, if so, label its location. Note that our ground truth data is quite different from those in existing datasets such as York Urban Dataset (YUD) [40] and Eurasian Cities Dataset (ECD) [8]. While these datasets are focused on urban scenes and attempt to identify *all* VPs in each image, our goal is to identify a *single dominant* VP associated with the main structures in a wide variety of scenes. The ability to identify the dominant VP in a scene is critical in our targeted applications related to aesthetics and photo composition.

Like existing datasets, we label the dominant VP by manually specifying at least two parallel lines in the image, denoted as $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$ (see Figure 1). The dominant VP location is then computed as $\boldsymbol{v} = \boldsymbol{l}_1 \times \boldsymbol{l}_2$. Because our goal is to identify the dominant VPs only, we make a few assumptions during the process. First, each VP must correspond to at least two visible parallel lines in the image. This eliminate other types of perspective in photography such as diminishing perspective, which is formed by placing identical or similar objects at different distances. Second, for a VP to be the dominant VP in an image, it must correspond to some major structures of the scene and clearly carries more visual weight than other candidates, if any. We do not consider images with two or more VPs carrying similar visual importance, which are typically seen in urban scenes. Similarly, we also exclude images where it is impossible to determine a single dominant direction due to parallel curves (Figure 2). Finally, observing that only those VPs which lie within or near the image frame convey a strong sense of perspective to the viewers, we resize each image so that the length of its longer side is 500 pixels, and only keep the dominant VPs that lie within a $1,000 \times 1,000$ frame, with the image placed at the center. We used the size 500 pixels as a reasonable compromise between keeping details and providing fast runtime for large-scale applications.

We collected a total of 1,316 images with annotations of ground truth parallel lines. The dataset is publicly available at `https://faculty.ist.psu.edu/zzhou/vp_labels.zip`.

Fig. 2. Example natural scene images that are **not** suitable for this work. The first two images show diminishing perspective. The third image has two VPs. The last image contains parallel curves, not parallel lines.

IV. Contour-based Vanishing Point Detection for Natural Scenes

Given a set of edges $\mathcal{E} = \{E_1, \ldots, E_N\}$, a VP detection method aims to classify the edges into several classes, one for each VP in the scene, plus an "outlier" class. Similar to [7], we employ the J-Linkage algorithm [14] for multiple model estimation and classification. The key new idea of our method lies in the use of contours to generate the input edges. As we will see in this section, our contour-based method can effectively identify weak edges in natural scene images and reduce the number of outliers at the same time, leading to significantly higher VP detection accuracy.

A. J-Linkage

Similar to RANSAC, J-Linkage first randomly chooses M minimal sample sets and computes a putative model for each of them. For VP detection, the j-th minimal set consists of two randomly chosen edges: (E_{j_1}, E_{j_2}). To this end, we first fit a line $\boldsymbol{l}_i$ to each edge $E_j \in \mathcal{E}$ using least squares. Then, we can generate the hypothesis $\boldsymbol{v}_j$ using the corresponding fitted lines: $\boldsymbol{v}_j = \boldsymbol{l}_{j_1} \times \boldsymbol{l}_{j_2}$.

Next, J-Linkage constructs a $N \times M$ preference matrix P, where the (i, j)-th entry is defined as:

$$p_{ij} = \begin{cases} 1 & \text{if } D(E_i, \boldsymbol{v}_j) \leq \phi \\ 0 & \text{otherwise} \end{cases} . \tag{1}$$

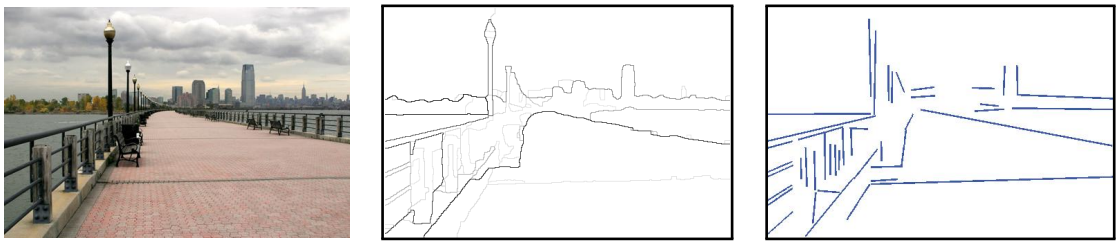
Here, $D(E_i, \boldsymbol{v}_j)$ is a measure of consistency between edge E_i and VP hypothesis $\boldsymbol{v}_j$, and ϕ is a threshold. Note that i-th row indicates the set of hypotheses edge E_i has given consensus to, and is called the *preference set* (PS) of E_i. J-Linkage then uses a bottom-up scheme to iteratively group edges that have similar PS. Here, the PS of a cluster is defined as the intersection of the preference sets of its members. In each iteration, the two clusters with the smallest distance are merged, where the Jaccard distance is used to measure the similarity between any two clusters A and B:

$$d_J(A, B) = \frac{|A \bigcup B| - |A \bigcap B|}{|A \bigcup B|} . \tag{2}$$

The operation is repeated until the distance between any two clusters is 1.

Consistency Measure. We intuitively define the consistency measure $D(E_i, \boldsymbol{v}_j)$ as the root mean square (RMS) distance from all points on E_i to a line $\hat{\boldsymbol{l}}$, such that $\hat{\boldsymbol{l}}$ passes through $\boldsymbol{v}_j$ and minimizes the distance:

$$D_{RMS}(E_i, \boldsymbol{v}_j) = \min_{l : l \times \boldsymbol{v}_j = 0} \left(\frac{1}{N} \sum_{\boldsymbol{p} \in E_i} dist(\boldsymbol{p}, \boldsymbol{l})^2 \right)^{\frac{1}{2}} , \tag{3}$$



(a) (b) (c)

Fig. 3. Contour-based edge detection. (a) Original image. (b) The ultrametric contour map (UCM). (c) Filtered edges.

where N is the number of points on E_i.

B. Edge Detection via Contours

Because we rely on edges to identify the dominant VP in an image, an ideal edge detection method should have the following properties: (i) it should detect all edges that converge to the true VPs, (ii) the detected edges should be as complete as possible, and (iii) it should keep the number of irrelevant or cluttered edges to a minimum. As we have discussed, local edge-detection methods do not meet these criteria. Instead, a successful method must go beyond local measurements and utilize *global* visual information.

Our key insight is that in order to determine if an edge is present at certain location, it is necessary to examine the *relevant regions* associated with it. This is motivated by the observation that humans label the edges by first identifying the physical objects in an image. In addition, based on the level of details they choose, different people may make different decisions on whether to label a particular edge.

Accordingly, for edge detection, we employ the widely-used contour detection method [41], which proposed a unified framework for contour detection and image segmentation using an agglomerative region clustering scheme. In the following, we first discuss the main difference between the contours and edges detected by local methods. Then we show how to obtain straight edges from the contours.

Globalization in Contour Detection. Comparing to the local methods, the contours detected by [41] enjoy two levels of globalization.

First, as a global formulation, *spectral clustering* has been widely used in image segmentation to suppress noise and boost weak edges. Generally, let W be an affinity matrix whose entries encode the (local) similarity between pixels, this method solves for the generalized eigenvectors of the linear system: $(D - W)\boldsymbol{v} = \lambda D\boldsymbol{v}$, where the diagonal matrix D is defined as $D_{ii} = \sum_j W_{ij}$. Let $\{\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}$ be the eigenvectors corresponding the $K+1$ smallest eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_K$. Using all the eigenvectors except $\boldsymbol{v}_0$, one can then represent each image pixel with a vector in $\mathbb{R}^K$. As shown in [41], the distances between these new vectors provide a denoised version of the original affinities, making them much easier to cluster.

Second, a graph-based *hierarchical clustering* algorithm is used in [41] to construct an *ultrametric contour map* (UCM) of the image (see Figure 3(b)). The UCM defines a duality between closed, non-self-intersecting weighted contours and a hierarchy of regions, where different levels of the hierarchy correspond to different levels of detail in the image. Thus, each weighted contour in UCM represents the dissimilarity of two, possibly large, regions in the image, rather than the local contrast of small patches.

From Contours to Edges. Let $\mathcal{C} = \{C_1, C_2, \ldots\}$ denote the set of all weighted contours. To recover straight edges from the contour map, we apply a scale-invariant contour subdivision procedure. Specifically, for any contour C_j, let $\boldsymbol{c}_j^1$ and $\boldsymbol{c}_j^2$ be the two endpoints of C_j, we first find the point on C_j which has the maximum distance to the straight line segment connecting its endpoints:

$$\boldsymbol{p}^* = \arg\max_{\boldsymbol{p} \in C_j} dist(\boldsymbol{p}, \overline{\boldsymbol{c}_j^1 \boldsymbol{c}_j^2}) \ . \tag{4}$$

We then subdivide C_j at $\boldsymbol{p}^*$ if the maximum distance is greater than a fixed fraction α of the contour length:

$$dist(\boldsymbol{p}^*, \overline{\boldsymbol{c}_j^1 \boldsymbol{c}_j^2}) > \alpha \cdot |C_j| \ . \tag{5}$$

By recursively applying the above procedure to all the contours, we obtain a set of approximately straight edges $\mathcal{E} = \{E_1, \ldots, E_N\}$. We only keep edges that are longer than certain threshold $l_{\min}$, because short edges are very sensitive to image noises (Figure 3(c)).

C. Experiments

In this section, we present a comprehensive performance study of our contour-based VP detection method, and compare it to the state-of-the-art. Similar to previous work (*e.g.*, [7], [9]), we evaluate the performance of a VP detection method based on the consistency of the ground truth edges with the estimated VPs. Specifically, let $\{E_k^G\}_{k=1}^K$ be the set of ground truth edges, the consistency error of a detection $\hat{\boldsymbol{v}}$ is:

$$err(\hat{\boldsymbol{v}}) = \frac{1}{K} \sum_k D_{RMS}(E_k^G, \hat{\boldsymbol{v}}) \ . \tag{6}$$

For all experiments, we compute the average consistency error over five independent trials.

1) Comparison of Edge Detection Methods: We first compare our contour-based edge detection to the popular Canny detector [12] and LSD [13] in terms of the accuracy of the detected VPs. For our contour-based method, the parameters are: $\alpha = 0.05$, $l_{\min} = 40$, and $\phi = 3$. For Canny detector and LSD, we tune the parameters $l_{\min}$ and ϕ so that the highest accuracy is obtained. In this experiment, we simply

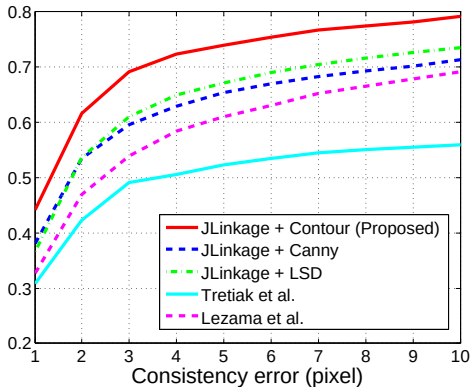


Fig. 4. Vanishing point detection results. We show the cumulative histograms of vanishing point consistency error w.r.t. the ground truth edges (Eq. (6)) for all candidate methods.

keep the VP with the largest support set as the detection result. Figure 4 reports the cumulative histograms of vanishing point consistency error w.r.t. the ground truth edges for all three methods. As one can see, our contour-based method significantly outperforms the other edge detection methods.

In Figure 5, we further show some example edge detection results. Note that, since most VP detection methods rely on clustering the detected edges, an ideal edge detector should maximize the number of edges consistent with the ground truth dominant VP, and minimize the number of irrelevant edges. As shown, our contour-based method can better detect weak yet important edges in terms of both the quantity and the completeness. For example, our method is able to detect the complete edges of the road in Figure 5(b), while the local methods only detected parts of them. Also, only our method successfully detected the edges of the road in Figure 5(c).

Another important distinction between our contour-based method and the local methods concerns the textured areas in the image. Local methods tend to confuse image texture with true edges, resulting a large number of detections in these areas (*e.g.*, the sky region and the road in 5(d) and (e), respectively). Such false positives often lead to incorrect clustering results in the subsequent VP detection stage. Meanwhile, our method treats the textured area as a whole, thus greatly reducing the number of false positives.

2) Comparison with the State-of-the-Art: Next, we compare our method to state-of-the-art VP detection methods. As we discussed before, most existing methods focus on urban scenes and make strong assumptions about the scene structures, such as a Manhattan world model [17], [18], [9], [19]. Such strong assumptions render these methods inapplicable to natural landscape scenes.

While other methods do not explicitly assume a specific model, they still benefit from the scene structures to various extents. In Figure 4, we compare our method to two recent methods, namely Tretiak *et al.* [8] and Lezama *et al.* [11]. Note that [11] uses the Number of False Alarms (NFA) to measure the importance of the detected VPs. For fair comparison, we keep the VP with the highest NFA. Figure 4 shows that the two methods do not perform well on the natural landscape images. The problem with [8] is that it assumes multiple horizontal VP detections for horizon and zenith estimation, but there may not be more than one VP in natural scenes. Similarly, [11] relies on the multiple horizontal VP detections to filter redundant and spurious VPs.

3) Parameter Sensitivity: We further study the performance of our contour-based VP detection method w.r.t. the parameters α, the minimum edge length $l_{\min}$, and the distance threshold ϕ in Eq. (1). We conduct experiments with one of these parameters varying while the others are fixed. The default parameter setting is $\alpha = 0.05$, $l_{\min} = 40$, and $\phi = 3$.

Performance w.r.t. α. Recall from Section IV-B that α controls the degree to which a contour segment may deviate from a straight line before it is divided into two sub-segments. Figure 6(a) shows that the best performance is achieved with $\alpha = 0.05$.

Performance w.r.t. minimum edge length $l_{\min}$. Figure 6(b) shows the performance of our method as a function of $l_{\min}$. Rather surprisingly, we find that the accuracy is quite sensitive to $l_{\min}$. This is probably because that, for natural scenes, the number of edges consistent with the dominant VP is relatively small. Therefore, if $l_{\min}$ is too small, these edges may be dominated by irrelevant edges in the scene; if $l_{\min}$ is too large, there may not be enough inliers to robustly estimate the VP location.

Performance w.r.t. threshold ϕ. Figure 6(c) shows the accuracy of our method w.r.t. the threshold ϕ in Eq. (1). As one can see, our method is relatively insensitive to the threshold, and achieves the best performance when $\phi = 3$.

V. Selection of the Dominant Vanishing Point

In real-world applications concerning natural scene photos, it is often necessary to select the images in which a dominant VP is present since many images do not have a VP. Further, if multiple VPs are detected, we need to determine which one carries the most importance in terms of the photo composition. Therefore, given a set of candidates $\{\boldsymbol{v}_j\}_{j=1}^n$ generated by a VP detection method, our goal is to find a function f which well estimates the strength of a VP candidate. Then, we can define the dominant VP of an image as the one whose strength is (i) the highest among all candidates, and (ii) higher than certain threshold T:

$$\boldsymbol{v}^* = \arg \max_{f(\boldsymbol{v}_j) \geq T} f(\boldsymbol{v}_j) \ . \tag{7}$$

In practice, given a detected VP $\boldsymbol{v}_j$ and the edges $\mathcal{E}_j \subseteq \mathcal{E}$ associated with the cluster obtained by a clustering method (*e.g.*, J-Linkage), a simple implementation of f would be the number of edges: $f(\boldsymbol{v}_j) = |\mathcal{E}_j|$. Note that it treats all edges in $\mathcal{E}_j$ equally. However, we have found that this is problematic for natural images because it does not consider the implied depth of each edge in the 3D space.

A. The Strength Measure

Intuitively, an edge conveys a strong sense of depth to the viewers if (i) it is long, and (ii) it is close to the VP (Figure 1). This observation motivates us to examine the implied depth of

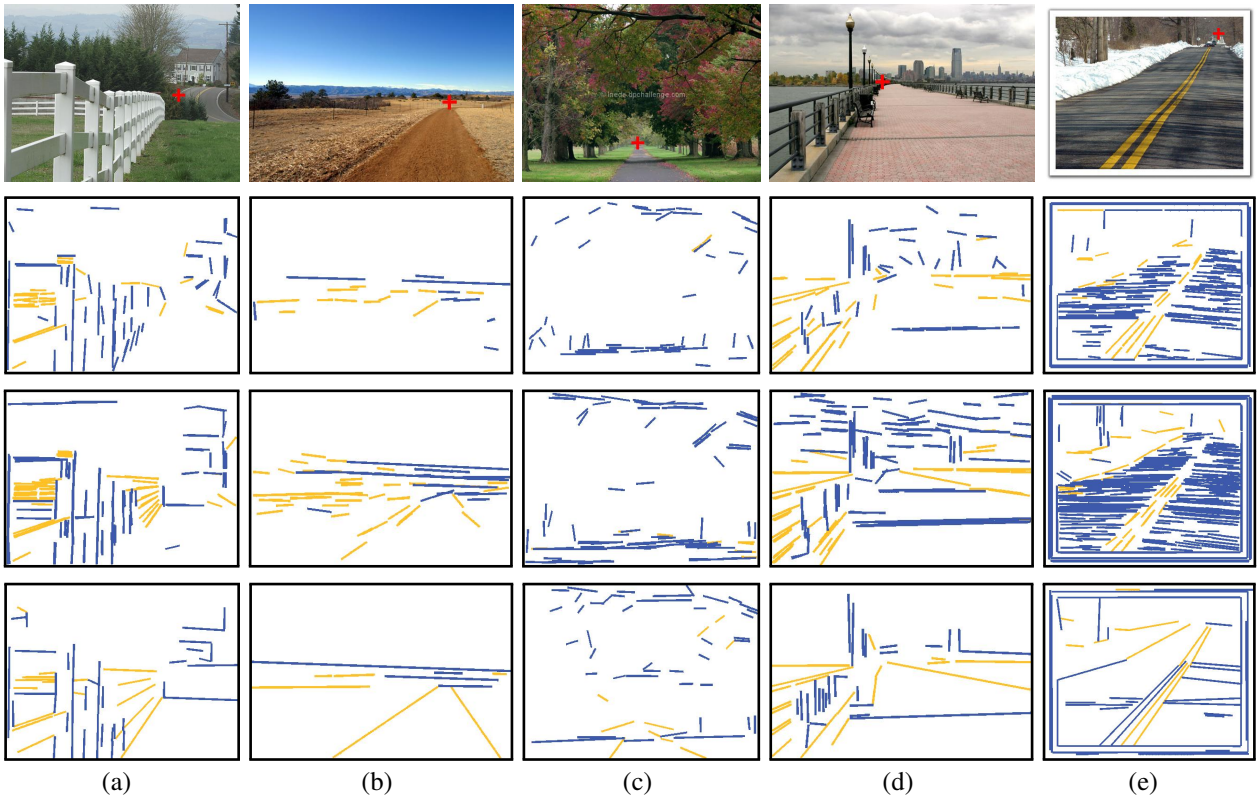


Fig. 5. Comparison of different edge detection methods. The four rows show the original images, and the edges detected by Canny detector, LSD, and our contour-based method, respectively. Yellow edges indicate the edges consistent with the ground truth dominant vanishing point.

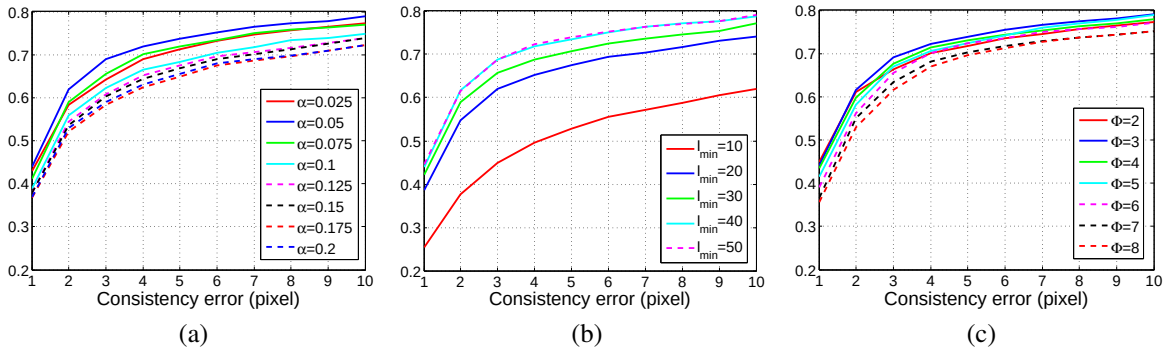


Fig. 6. Accuracy of our method w.r.t. parameters α, $l_{\min}$, and ϕ, respectively.

each individual point on an edge, instead of treating the edge as a whole.

Geometrically, as shown in Figure 7, let E be a line segment consistent with vanishing point $\boldsymbol{v} = (v_x, v_y, 1)^T$ in the image.[1] We further let D be the direction in 3D space (*i.e.*, a point at infinity) that corresponds to $\boldsymbol{v}$: $\boldsymbol{v} = PD$, where $P \in \mathbb{R}^{3\times4}$ is the camera projection matrix.

For any pixel on the line segment $\boldsymbol{q} = (q_x, q_y, 1)^T \in E$, we denote Q as the corresponding point in the 3D space. Then, we can represent Q as a point on a 3D line with direction D: $Q = A + \lambda D$, where A is some reference point chosen on this line, and λ can be regarded as the (relative) distance between

[1]In this section, all 2D and 3D points are represented in homogeneous coordinates.

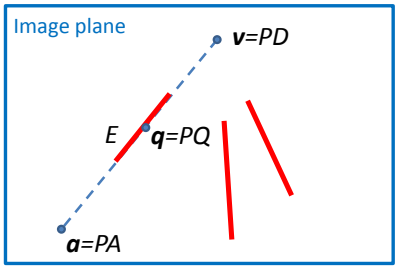


Fig. 7. Illustration of our edge strength measure.

A and Q. Consequently, we have

$$\boldsymbol{q} = PQ = P(A + \lambda D) = \boldsymbol{a} + \lambda \boldsymbol{v} \ , \tag{8}$$

where $\boldsymbol{a} = (a_x, a_y, 1)^T$ is the image of A. Thus, let $l_{\boldsymbol{q}}$ and

$l_{\boldsymbol{a}}$ denote the distance on the image from $\boldsymbol{q}$ and $\boldsymbol{a}$ to $\boldsymbol{v}$, respectively, we have

$$\lambda = l_{\boldsymbol{a}}/l_{\boldsymbol{q}} - 1 \ . \tag{9}$$

Note that if we choose A as the intersecting point of the 3D line corresponding to E and the image plane, λ represents the (relative) distance from any point Q on this line to the image plane along direction D. In practice, although $l_{\boldsymbol{a}}$ is typically unknown and varies for each edge E, we can still infer from Eq. (9) that λ is a linear function of $1/l_{\boldsymbol{q}}$. This motivates us to define the weight of a pixel $\boldsymbol{q} \in E$ as $(l_{\boldsymbol{q}} + \tau)^{-1}$, where τ is a constant chosen to make it robust to noises and outliers.

Thus, our new measure of strength for $\boldsymbol{v}_j$ is defined as

$$f(\boldsymbol{v}_j) = \sum_{E \in \mathcal{E}_j} \sum_{\boldsymbol{q} \in E} \frac{1}{l_{\boldsymbol{q}} + \tau} \ . \tag{10}$$

Clearly, edges that are longer and closer to the VP have more weights according to our new measure.

B. Experiments

1) Dominant Vanishing Point Selection: We first demonstrate the effectiveness of the proposed strength measure in selecting the dominant VP from the candidates obtained by our VP detection algorithm. In Figure 8, we compare the following three measures in terms of the consistency error of the selected dominant VP:

`Edge Num`: The number of edges associated with each VP.

`Edge Sum`: The sum of the edge lengths associated with each VP.

`Proposed`: Our strength measure Eq. (10).

As shown, by considering the length of an edge and its proximity to the VP, our proposed measure achieves the best performance in selecting the dominant VP in the image.

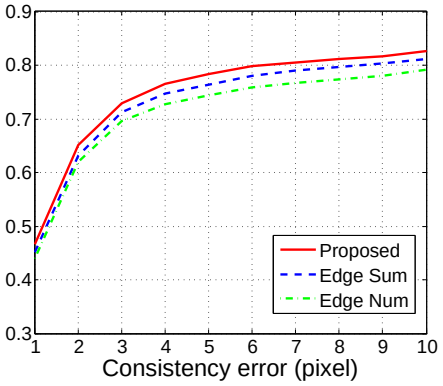


Fig. 8. Experiment results on dominant vanishing point selection.

2) Dominant Vanishing Point Verification: Next, we evaluate the effectiveness of the proposed measure in determining the existence of a dominant VP in the image. For this experiment, we use all the 1,316 images with labeled dominant VPs as positive samples, and randomly select 1,500 images without a VP from the "landscape" category of AVA dataset as negative samples. In Figure 9, we plot the ROC curves of the three different measures. As a baseline, we also include the result of the Number of False Alarms (NFA) score proposed in [11], which measures the likelihood that a specific configuration (*i.e.*, a VP) arises from a random image. One can clearly see that our proposed measure achieves the best performance.

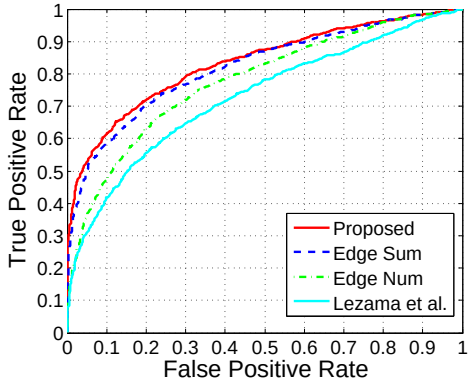
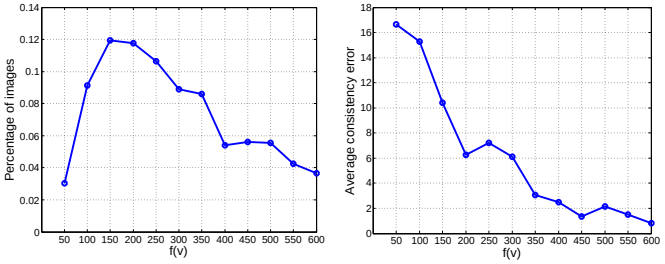


Fig. 9. Experiment results on dominant vanishing point verification.

In Figure 10(a) and (b), we further plot the percentage of images as a function of our strength measure, and the average consistency error, respectively. In particular, Figure 10(b) shows that the consistency error decreases substantially when the strength score is higher than 150. This suggests that our strength measure is a good indicator of the reliability of a VP detection result.



(a) (b)

Fig. 10. The impact of VP strength on the accuracy of dominant VP detection. We show (a) the percentage of images and (b) the average consistency error as a function of our strength measure Eq. (10).

VI. Photo Composition Application

The dominant VP and the image elements associated with it (*e.g.*, parallel lines, planes) encode rich information about the 3D scene geometry. Such information enables us to gain deeper understanding of the perspective effect in the photos, hence can potentially facilitate many tasks in photography. In this section, we demonstrate an interesting application in automatic understanding of photo composition that aims to provide on-site guidance to amateur photographers on their work through composition-sensitive image retrieval.

Cloud-based photo sharing services such as `flickr.com`, `photo.net`, and `dpchallenge.com` allow photographers to access millions of photos taken by their peers around the world. Such resources have been playing an increasingly

important role in helping the amateurs improve their photography skills. Specifically, considering the scenario where a photographer is about to the take a photo of a natural scene, he or she may wonder what photos peers or professional photographers would take in a similar situation. Therefore, given a shot taken by the user, we propose to find exemplar photos about *similar scenes* with *similar points of view* in a large collection of photos. These photos can then be used as feedback to the user on his or her own work.

Meanwhile, as we illustrate in Figure 1, professional photographers and visual artists have long realized that linear perspective provides us strong cues about the viewer's position and angle of perception. Motivated by this key insight, we propose a novel similarity measure based on the detected dominant VP in the image for *viewpoint-specific image retrieval*. Note that the idea of exploiting information about VPs for retrieval has been previously studied in [38]. But their method assumes known VP locations in *all* images, hence cannot readily be applied to general image collections where the majority may not contain a VP. In contrast, we use our new VP detection method and strength measure to automatically detect the dominant VPs in the images. Further, [38] does not consider image semantics for retrieval, hence its usage can be limited in practice.

A. Viewpoint-Specific Image Retrieval

Given two images I_i and I_j, our similarity measure is a sum of two components:

$$D(I_i, I_j) = D_s(I_i, I_j) + D_p(I_i, I_j), \tag{11}$$

where D_s and D_p measures the similarity of two images in terms of the scene semantics and the use of linear perspective, respectively. Below we describe each term in detail.

Semantic Similarity D_s: Recently, it has been shown that generic descriptors extracted from the convolutional neural networks (CNNs) are powerful in capturing the image semantics (*e.g.*, scene types, objects), and have been successfully applied to obtain state-of-the-art image retrieval results [37]. In our experiment, we adopt the publicly available CNN model trained by [42] on the ImageNet ILSVRC challenge dataset[2] to compute the semantic similarity. Specifically, we represent each image using the ℓ_2-normalized output of the second fully connected layer (full7 of [42]), and adopt the cosine distance to measure the feature similarity.

Perspective Similarity D_p: To model the perspective effect in the image, we consider two main factors: (i) the location of the dominant VP and (ii) the position of the associated image elements. For the latter, we focus on the edges consistent with the dominant VP obtained via our contour-based VP detection algorithm. Let $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ be the locations of the dominant VPs in images I_i and I_j, respectively. We use $\mathcal{E}_i$ (or $\mathcal{E}_j$) to denote the sets of edges consistent with $\boldsymbol{v}_i$ (or $\boldsymbol{v}_j$). Our perspective similarity measure is defined as:

$$D_p(I_i, I_j) = \gamma_1 \max\left(1 - \frac{\|\boldsymbol{v}_i - \boldsymbol{v}_j\|}{len}, 0\right) + \gamma_2 K(\mathcal{E}_i, \mathcal{E}_j)\ , \tag{12}$$

[2]http://www.vlfeat.org/matconvnet/pretrained/

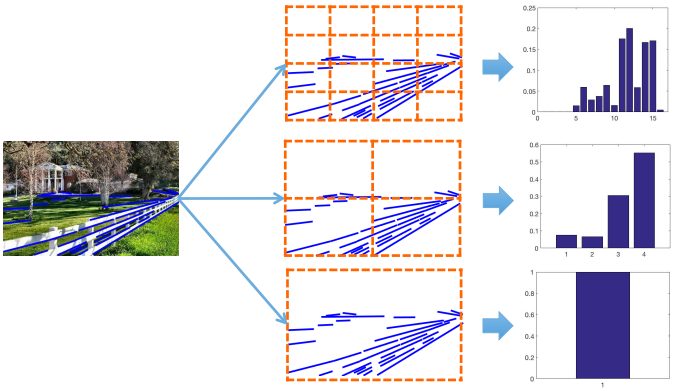


Fig. 11. Illustration of the construction of a three-level spatial pyramid. Given the set of edges $\mathcal{E}_i$ consistent with the dominant VP, we subdivide the image at three different level of resolutions. For each resolution, we count the number of edge points that fall into each bin to build the histograms $H_i^l, l = 0, 1, 2$.

where $\|\boldsymbol{v}_i - \boldsymbol{v}_j\|$ is the Euclidean distance between $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$, len is the length of the longer side of the image. We resize all the images to $len = 500$.

Further, since each edge can be regarded as a set of 2D points on the image, $K(\mathcal{E}_i, \mathcal{E}_j)$ should measure the similarity of two point sets. Here, we use the popular *spatial pyramid matching* [43] for its simplicity and efficiency. Generally speaking, this matching scheme is based on a series of increasingly coarser grids on the image. At any fixed resolution, two points are said to match if they fall into the same grid cell. The final matching score is a weighted sum of the number of matches that occur at each level of resolution, where matches found at finer resolutions have higher weights than matches found at coarser resolutions.

For our problem, we first construct a series of grids at resolutions $0, 1, \ldots, L$, as illustrated in Figure 11. Note that the grid at level l has 2^l cells along each dimension, so the total number of cells is 2^{2l}. At level l, let $H_i^l(k)$ and $H_j^l(k)$ denote the number of points from $\mathcal{E}_i$ and $\mathcal{E}_j$ that fall into the k-th cell, respectively. The number of matches at level l is then given by the histogram intersection function:

$$\mathcal{I}(H_i^l, H_j^l) = \sum_k \min\left(H_i^l(k), H_j^l(k)\right)\ . \tag{13}$$

Below we write $\mathcal{I}(H_i^l, H_j^l)$ as $\mathcal{I}^l$ for short.

Since the number of matches found at level l also includes all the matches found at the level $l+1$, the number of new matches at level l is given by $\mathcal{I}^l - \mathcal{I}^{l+1}$, $\forall l = 0, \ldots, l-1$. To reward matches found at finer levels, we assign weight $2^{-(L-l)}$ to the matches at level l. Note that the weight is inversely proportional to the cell width at that level. Finally, the pyramid matching score is defined as:

$$K^L(\mathcal{E}_i, \mathcal{E}_j) = \mathcal{I}^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}}(\mathcal{I}^l - \mathcal{I}^{l+1}) \tag{14}$$

$$= \frac{1}{2^L}\mathcal{I}^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}}\mathcal{I}^l\ . \tag{15}$$

Here, we use superscript "L" to indicate its dependency on the parameter L. We empirically set the parameters for viewpoint-

specific image retrieval for all experiments to: $\gamma_1 = \gamma_2 = 0.5, L = 6$.

B. Case Studies

In our experiments, we use the entire "landscape" category of the AVA dataset to study the effectiveness of our new similarity measure. We first run our contour-based VP detection algorithm to detect the dominant VP in the 21,982 images in that category. We only keep those VPs with strength scores higher than 150 for this study because, as we discussed in Section V-B2, detections with low strength scores are often unreliable. If no dominant VP is detected in an image, we simply set the perspective similarity $D_p(I_i, I_j) = 0$.

Figure 12 shows the top-ranked images for various query images in the AVA dataset. It is clear that our method is able to retrieve images with similar content *and* similar viewpoints as the query image. More importantly, the retrieved images exhibit a wide variety in terms of the photographic techniques used, including color, lighting, photographic elements, design principles, etc. Thus, by examining the exemplar images retrieved by our system, amateur photographers may conveniently learn useful techniques to improve the quality of their work. Below we examine a few cases:

1^{st} row, red boxes: This pair of photos highlight the importance of lighting in photography. Instead of taking the picture with overcast, it is better to take the shot close to sunset with clear sky, as the sunlight and shadows could make the photo more attractive. Also, it can be desirable to leave more water in the scene.

6^{th} row, blue boxes: These three photos illustrate the different choices of framing and photo composition. While the query image uses a diagonal composition, alternative ways to shoot the bridge scene include using a vertical frame or lowering the camera to include the river.

9^{th} row, green boxes: This case shows an example where photographers sometimes choose unconventional aspect ratios (*e.g.*, a wide view) to make the photo more interesting.

11^{th} row, yellow boxes: Compared to the query image, the retrieved image contains more vivid colors (the grass) and texture (the cloud).

Besides, the last two rows of Figure 12 show the typically failure cases of our method, in which we also plot the edges correspond to the detect VP in the query image. In the first example, our VP detection method fails to detect the true dominant VP in the image. For the second example, while the detection is successful, our retrieval system is unable to find images with similar viewpoints. In real-world applications, however, we expect the size of the image database to be much larger than our experimental database (with $\sim$20K images) and the algorithm should be able to retrieve valid results.

C. Comparison to the State-of-the-Art

We compare our method to two popular retrieval systems, which are based on the HOG features [44], [45] and the CNN features, respectively. While many image retrieval methods exist in the literature, we choose the two because (i) the CNN features have been recently shown to achieve state-of-the-art performance on semantic image retrieval; and (ii) similar to our method, the HOG features are known to be sensitive to image edges, thus serve as a good baseline for comparison.

For **HOG**, We represent each image with a rigid grid-like HOG feature $\boldsymbol{x}_i$ [44], [45]. As suggested in [36], we limit its dimensionality to roughly $5K$ by resizing the images to 150×100 or 100×150 and using a cell size of 8 pixels. The feature vectors are normalized by subtracting the mean: $\boldsymbol{x}_i = \boldsymbol{x}_i - mean(\boldsymbol{x}_i)$. We use the cosine distance as the similarity measure. For **CNN**, we directly use $D_s(I_j, I_j)$ discussed in Section VI-A as the final matching score. Obviously, our method reduces to **CNN** if we set $\gamma_1 = \gamma_2 = 0$ in Eq. (12).

Figure 13 shows the best matches retrieved by all systems for various query images. Both **CNN** and our method are able to retrieve semantically relevant images. However, the images retrieved by **CNN** vary significantly in terms of the viewpoint. In contrast, out method is able to retrieve images with similar viewpoints. While **HOG** is somewhat sensitive to the edge orientations (see the first and fourth examples in Figure 13), it is not as effective compared to our method in capturing the viewpoints and perspective effects.

Quantitative Human Subject Evaluation. We further perform a quantitative evaluation on the performance of our retrieval method. Unlike traditional image retrieval benchmarks, currently there is no dataset with ground truth composition (*i.e.*, viewpoint) labels available. In view of this barrier, we have instead conducted a user study which asks participants to manually compare the performance of our method with that of **CNN** based on their ability to retrieve images that have *similar semantics and similar viewpoints*. Note that we have excluded **HOG** from this study because (i) it performs substantially worse than our method and **CNN**, and (ii) we are particularly interested in the effectiveness of the new perspective similarity measure D_p.

In this study, a collection of 200 query images (with VP strength scores higher than 150) are randomly selected from our new dataset of 1,316 images that each containing a dominant VP (Section III). At our user study website, each participant is assigned with a subset of 30 randomly selected query images. For each query, we show the top-8 images retrieved by both systems and ask the participant to rank the performance of the two systems. To avoid any biases, no information about the two systems was provided during the study. Further, we randomly shuffled the order in which the results of the two systems are shown on each page.

We have recruited 10 participants to this study, mostly graduate students with some basic photography knowledge. Overall, our system is ranked better for 76.7% of the time, whereas **CNN** is ranked better for only 23.3% of the time. This suggests our system significantly outperforms the state-of-the-art for the viewpoint-specific image retrieval task.

VII. Conclusions

In this paper, we study an intriguing problem of detecting vanishing points in natural landscape images. We develop a new VP detection method, which combines a contour-based

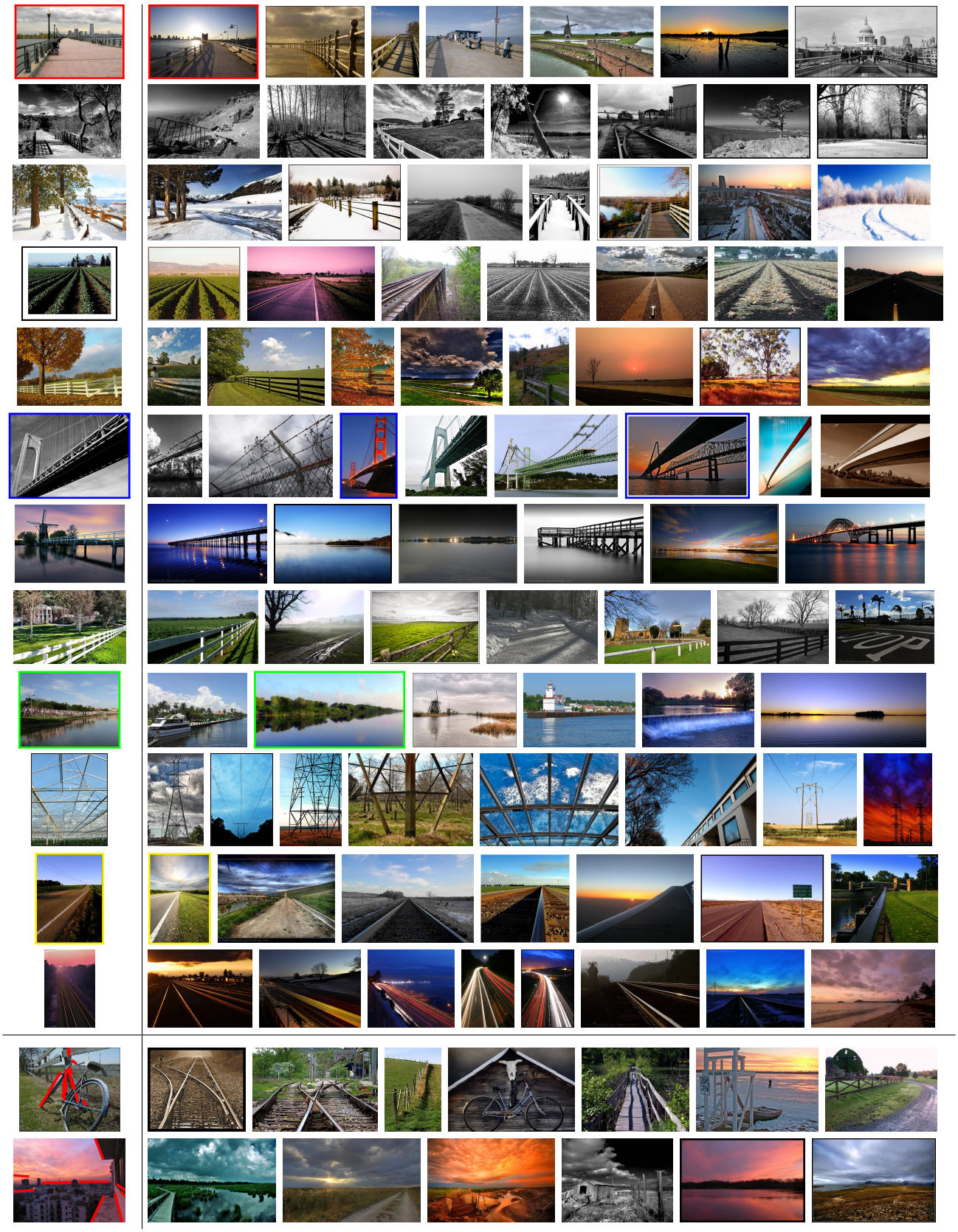


Fig. 12. Viewpoint-specific image retrieval results. Each row shows a query image (first image from the left) and the top-ranked images retrieved by our method. Last two rows show some failure cases.

Fig. 13. Comparison to state-of-the-art retrieval methods. For a query image, we show the top four or five images retrieved by different methods, where each row corresponds to one method. **First row:** Our method. **Second row:** CNN. **Third row:** HOG.

edge detector with J-Linkage, and show that it outperforms state-of-the-art methods on a new ground truth dataset. The detected VPs and the associated image elements provides valuable information about the photo composition. As an application of our method, we develop a novel viewpoint-specific image retrieval system which can potentially provide useful on-site feedback to photographers.

One limitation of our current system is that it is not designed to handle images in which the linear perspective is absent. For example, to convey a sense of depth, other techniques such as diminishing objects and atmospheric perspective have also been used. Meanwhile, instead of relying solely on the linear perspective, experienced photographers often employ multiple design principles such as balance, contrast, unity, and illumination. In the future, we plan to explore these factors for extensive understanding of photo composition.

REFERENCES

[1] B. P. Krages, *Photography: The Art of Composition*. Allworth Press, 2005.

[2] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *ACM Multimedia*, 2010, pp. 271–280.

[3] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or, "Optimizing photo composition," *Comput. Graph. Forum*, vol. 29, no. 2, pp. 469–478, 2010.

[4] L. Yao, P. Suryanarayan, M. Qiao, J. Z. Wang, and J. Li, "Oscar: On-site composition and aesthetics feedback through exemplars for photographers," *International Journal of Computer Vision*, vol. 96, no. 3, pp. 353–383, 2012.

[5] B. Ni, M. Xu, B. Cheng, M. Wang, S. Yan, and Q. Tian, "Learning to photograph: A compositional perspective," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1138–1151, 2013. [Online]. Available: http://dx.doi.org/10.1109/TMM.2013.2241042

[6] J. Kosecká and W. Zhang, "Video compass," in *ECCV (4)*, 2002, pp. 476–490.

[7] J.-P. Tardif, "Non-iterative approach for fast and accurate vanishing point detection," in *ICCV*, 2009, pp. 1250–1257.

[8] E. Tretiak, O. Barinova, P. Kohli, and V. S. Lempitsky, "Geometric image parsing in man-made environments," *International Journal of Computer Vision*, vol. 97, no. 3, pp. 305–321, 2012.

[9] H. Wildenauer and A. Hanbury, "Robust camera self-calibration from monocular images of manhattan worlds," in *CVPR*, 2012, pp. 2831–2838.

[10] Y. Xu, S. Oh, and A. Hoogs, "A minimum error vanishing point detection approach for uncalibrated monocular images of man-made environments," in *CVPR*, 2013, pp. 1376–1383.

[11] J. Lezama, R. G. von Gioi, G. Randall, and J. Morel, "Finding vanishing points via point alignments in image primal and dual domains," in *CVPR*, 2014, pp. 509–515.

[12] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, 1986.

[13] R. G. von Gioi, J. Jakubowicz, J. Morel, and G. Randall, "LSD: A fast line segment detector with a false detection control," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 722–732, 2010.

[14] R. Toldo and A. Fusiello, "Robust multiple structures estimation with j-linkage," in *ECCV (1)*, 2008, pp. 537–547.

[15] D. A. Lauer and S. Pentak, *Design Basics*. Cengage Learning, 2011.

[16] A. Vedaldi and A. Zisserman, "Self-similar sketch," in *ECCV*, 2012, pp. 87–100.

[17] F. M. Mirzaei and S. I. Roumeliotis, "Optimal estimation of vanishing points in a manhattan world," in *ICCV*, 2011, pp. 2454–2461.

[18] J. C. Bazin, Y. Seo, C. Demonceaux, P. Vasseur, K. Ikeuchi, I. Kweon, and M. Pollefeys, "Globally optimal line clustering and vanishing point estimation in manhattan world," in *CVPR*, 2012, pp. 638–645.

[19] M. Antunes and J. P. Barreto, "A global approach for the detection of vanishing points and mutually orthogonal vanishing directions," in *CVPR*, 2013, pp. 1336–1343.

[20] C. Rasmussen, "Grouping dominant orientations for ill-structured road following," in *CVPR (1)*, 2004, pp. 470–477.

[21] H. Kong, J.-Y. Audibert, and J. Ponce, "Vanishing point detection for road detection," in *CVPR*, 2009, pp. 96–103.

[22] P. Moghadam, J. A. Starzyk, and W. S. Wijesoma, "Fast vanishing-point detection in unstructured environments," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 425–430, 2012.

[23] F. Zhang, M. Wang, and S. Hu, "Aesthetic image enhancement by dependence-aware object recomposition," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1480–1490, 2013.

[24] C. Fang, Z. Lin, R. Mech, and X. Shen, "Automatic image cropping using visual composition, boundary simplicity and content preservation models," in *ACM Multimedia*, 2014, pp. 1105–1108.

[25] M. B. Islam, W. Lai-Kuan, and W. Chee-Onn, "A survey of aesthetics-driven image recomposition," *Multimedia Tools and Applications*, pp. 1–26, 2016.
[26] J. Yan, S. Lin, S. B. Kang, and X. Tang, "Change-based image cropping with exclusion and compositional features," *International Journal of Computer Vision*, vol. 114, no. 1, pp. 74–87, 2015.
[27] R. Hong, L. Zhang, and D. Tao, "Unified photo enhancement by discovering aesthetic communities from flickr," *IEEE Trans. Image Processing*, vol. 25, no. 3, pp. 1124–1135, 2016.
[28] H. Su, T. Chen, C. Kao, W. H. Hsu, and S. Chien, "Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features," *IEEE Trans. Multimedia*, vol. 14, no. 3-2, pp. 833–843, 2012.
[29] S. Ma, Y. Fan, and C. W. Chen, "Pose maker: A pose recommendation system for person in the landscape photographing," in *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, 2014, pp. 1053–1056.
[30] P. Xu, H. Yao, R. Ji, X. Liu, and X. Sun, "Where should I stand? learning based human position recommendation for mobile photographing," *Multimedia Tools Appl.*, vol. 69, no. 1, pp. 3–29, 2014.
[31] Y. S. Rawat and M. S. Kankanhalli, "Context-aware photography learning for smart mobile devices," *TOMCCAP*, vol. 12, no. 1s, p. 19, 2015.
[32] W. Yin, T. Mei, C. W. Chen, and S. Li, "Socialized mobile photography: Learning to photograph with social context via mobile devices," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 184–200, 2014.
[33] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1349–1380, 2000.
[34] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 5:1–60, 2008.
[35] H. Jegou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010, pp. 3304–3311.
[36] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros, "Data-driven visual similarity for cross-domain image matching," *ACM Transactions on Graphics*, vol. 30, no. 6, p. 154, 2011.
[37] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *CVPR Workshops*, 2014, pp. 512–519.
[38] Z. Zhou, S. He, J. Li, and J. Z. Wang, "Modeling perspective effects in photographic composition," in *ACM Multimedia*, 2015, pp. 301–310.
[39] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *CVPR*, 2012, pp. 2408–2415.
[40] P. Denis, J. H. Elder, and F. J. Estrada, "Efficient edge-based methods for estimating manhattan frames in urban imagery," in *ECCV*, 2008, pp. 197–210.
[41] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, 2011.
[42] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference*, 2014.
[43] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006, pp. 2169–2178.
[44] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, pp. 886–893.
[45] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.