# A VANISHING POINT-BASED GLOBAL DESCRIPTOR FOR MANHATTAN SCENES

*Rohit Naini*[*]

University of Illinois
Urbana, IL 61801, USA.

*Shantanu Rane, Srikumar Ramalingam*

Mitsubishi Electric Research Laboratories
Cambridge, MA 02139, USA.

## ABSTRACT

Viewpoint-invariant object matching is challenging due to image distortions caused by several factors such as rotation, translation, illumination, cropping and occlusion. We propose a compact, global image descriptor for Manhattan scenes that captures relative locations and strengths of edges along vanishing directions. To construct the descriptor, an edge map is determined per vanishing point, capturing the edge strengths over a range of angles measured at the vanishing point. For matching, descriptors from two scenes are compared across multiple candidate scales and displacements. The matching performance is refined by comparing edge shapes at the local maxima of the scale-displacement plots. The proposed descriptor matching algorithm achieves an equal error rate of 7% for the Zurich Buildings Database, indicating significant gains in discriminative ability over other global descriptors that rely on aggregate image statistics but do not exploit the underlying scene geometry.

***Index Terms***— Global descriptors, vanishing points, Manhattan scenes

## 1. INTRODUCTION

Visual scene understanding is a long-standing open problem in computer vision. In particular, identification of objects in a 3D scene based on their projection onto a 2D image plane poses formidable challenges. The visual cortex is known to rely heavily on the presence of edges at physical object boundaries for identifying individual objects within a view [1]. Using cues from edges, texture and color, the brain is usually able to visualize and understand a 3D scene irrespective of the observer's viewpoint. In contrast, lacking a high level processing architecture like the visual cortex, modern computers must explicitly incorporate low-level viewpoint invariance into scene descriptors.

Approaches to scene understanding in the literature can be divided into two broad classes. One class relies on local interest points —also referred to as *keypoints*— that are robustly detected irrespective of rotation, translation and other viewpoint changes. A descriptor is then constructed around the keypoint so as to capture the local structure of gradients, texture, color and other information which remains invariant to viewpoint changes. SIFT [2] and SURF [3] are just two out of a growing palette of such keypoint-based descriptors. Another class of methods involves capturing features at a global scope and introducing robustness by local averaging and by using other statistical properties of color and gradient distributions. This global approach is employed in HOG [4] and GIST [5] descriptors.

The local and global approaches have complementary features. Local descriptors are highly robust and fairly discriminative for the corresponding keypoint, but global structural cues about the larger object in question are absent and can only be inferred after establishing correspondences among several local descriptors associated with the key points [6] [7]. Global descriptors tend to capture aggregate statistical information about the image but again, do not include specific geometric or structural cues which are often relevant for scene understanding. In this work, we propose a scene/object descriptor that not only has a global scope but also retains useful geometric and spatial information corresponding to image edges.

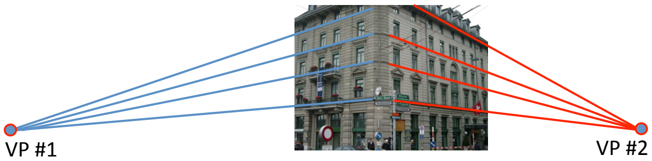

**Fig. 1**: 2 vanishing points of a Manhattan scene. There is a third VP above the image, which is not shown.

Our approach is limited to scenarios where the underlying scenes or objects have dominant directional orientations usually (not necessarily) in 3 orthogonal directions. These are commonly referred to in the literature as Manhattan scenes or geometry. Many man-made scenes satisfy the Manhattan world assumption [8], where the lines are oriented along three principal orthogonal directions. A crucial aspect of Manhattan geometry is that all parallel lines with a dominant direction intersect at a *vanishing point* in the 2D image plane, as depicted in Fig. 1. In scenes where three orthogonal directions may not exist, they may satisfy a single dominant direction (vertical) as in "Atlanta World" [9], or contain multiple dominant non-orthogonal directions as in items of furniture. Different from prior global descriptors, we exploit these geometrical constraints to build a compact and discriminative descriptor for Manhattan scenes. The ensuing development consists of a recipe for the construction of the descriptor (Section 2), an algorithm for matching descriptors to find similar objects (Section 3), and a description of our experiments of object matching on a public-domain database (Section 4).

## 2. VANISHING POINT-BASED IMAGE DESCRIPTOR

The proposed descriptor is based on the following two observations about multiple images (views) of the same object. First, parallel lines strictly maintain their angular ordering across images (up to an inversion) when they intersect at a vanishing point. Second, the relative lengths and relative angles of the parallel lines meeting at a vanishing point are approximately the same. These observations suggest that the relative locations and strengths of edges oriented along the vanishing directions can be used to build a descriptor. We describe the steps involved in constructing the descriptor below.

[*]This work was carried out when R. Naini was an intern at MERL.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

### 2.1. Seeding Descriptors at each Vanishing Point

A vanishing point (VP) is defined as the point of intersection of projections of lines which are parallel in the 3D scene. A VP can be considered as the 2D projection of a 3D point infinitely far away in the direction given by parallel lines in the 3D scene. In general, there are many vanishing points corresponding to multiple scene directions determined by parallel lines. Many man-made structures, e.g., urban landscapes, however have a regular cuboid geometry. Hence, usually, three vanishing points result from an image projection, 2 of which are shown in Fig. 1. VPs have been used in computer vision for image rectification, camera calibration [10] and related problems. Identification of VPs is simple if parallel lines in the underlying 3D scene are labeled, but becomes more difficult when labeling is not available. Methods for determining vanishing points include agglomerative clustering of edges [11], 1D Hough transforms [12], multi-level RANSAC-based approaches [13] and Expectation Maximization (EM) for assigning edges to VPs [14].

Denote the VP locations by $\overline{v}_i = (v_{ix}, v_{iy})$, $1 \leq i \leq m$. Typically for Manhattan scenes $m \leq 3$. Further, let $\theta_j(x, y)$ be the angle subtended at the VP $\overline{v}_j$ with respect to a reference line (horizontal) as shown in Fig. 2. Thus, $\theta_j(x, y) = \tan^{-1}(\frac{y - v_{jy}}{x - v_{jx}})$. The proposed descriptor is constructed by encoding the relative locations and strengths of the edges that converge at each VP. Thus, the descriptor can be considered as a function $D : \Theta \to \mathbf{R}^+$, whose domain consists of angular orientations of the edges converging at the VP under consideration, and whose range consists of some measure of the strengths of the corresponding edges. A descriptor is determined for each VP according to the process described below.

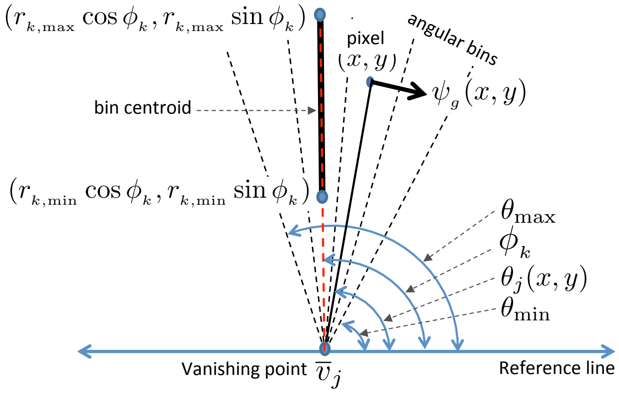
$(r_{k,\max} \cos \phi_k, r_{k,\max} \sin \phi_k)$
angular bins
pixel
$(x, y)$
$\psi_g(x, y)$
bin centroid
$(r_{k,\min} \cos \phi_k, r_{k,\min} \sin \phi_k)$
$\theta_{\max}$
$\phi_k$
$\theta_j(x, y)$
$\theta_{\min}$
Vanishing point $\overline{v}_j$
Reference line

**Fig. 2**: Uniform angular binning, various angles & gradients.

### 2.2. Edge Location Encoding

Line detection algorithms often produce broken and cropped lines, miss important edges and produce spurious ones. Therefore, for improved robustness, we propose to directly work with raw edge pixels, rather than lines that are fitted to image edges. Specifically, our goal is to bin the pixels whose gradients indicate that they are oriented according to the vanishing points for building a descriptor. To do this, we first compute the gradient $g(x, y)$, a 2D vector for every pixel in the image. The direction of a gradient of a pixel at $(x, y)$ refers to the direction along which there is a steep intensity variation. The magnitude of the gradient refers to the intensity difference at that pixel along the gradient direction. Let $\psi_g(x, y)$ and $|g(x, y)|$ refer to the direction and magnitude of the gradient vector $g(x, y)$. Then, compute a candidate pixel set $\mathcal{P}_j$ for the VP $\overline{v}_j$ as

$$\mathcal{P}_j = \left\{ (x, y) \middle| \left| \psi_g(x, y) - \theta_j(x, y) - \frac{\pi}{2} \right| \leq \tau \right\}$$

where $\tau$ is a threshold selected based on the amount by which the gradient direction is misaligned with the VP direction. Having determined $\mathcal{P}_j$, the underlying edge locations are encoded as follows: The pixel angles are quantized into a preset number $(K)$ of uniform angular bins centered at $\phi_k$, $1 \leq k \leq K$ within the angular range $[\theta_{\min}, \theta_{\max}]$ spanning the image from the VP, such that

$$\phi_k = \theta_{\min} + \frac{k}{K+1} \left( \theta_{\max} - \theta_{\min} \right), \quad 1 \leq k \leq K$$

### 2.3. Edge Strength Encoding

Studies on the human visual system suggest that the relative prominence of edges plays a role in visualizing a distinctive object pattern. The prominence of an image edge is a function of the length of the edge, its thickness, and the lateral variation (intensity and fall-off characteristics) in the direction perpendicular to the edge. There are several ways to construct an edge strength metric. For example, if edge detectors are used to construct the descriptor for a particular VP, the strength could be a function of the edge length and the pixel-wise cumulative gradient along the edge. However, as mentioned earlier, using edge detectors is not robust, so we prefer methods based on clustering or quantization of pixel-wise gradients. The process is described in detail below.

When the pixel set $\mathcal{P}_j$ is uniformly quantized into angular bins, one way to capture the edge strength is to compute the sum of the magnitudes of the gradients $|g(x, y)|$, in each quantization bin. To achieve this, we first consider a line segment passing through the middle of every angular quantization bin with adaptive end points $(r_{k,\min} \cos \phi_k, r_{k,\min} \sin \phi_k)$ and $(r_{k,\max} \cos \phi_k, r_{k,\max} \sin \phi_k)$, with $r_{k,\min}$ and $r_{k,\max}$ spanning the extent of the bin as shown in Fig. 2. The descriptor is then given by:

$$D(k) = \sum_{r=r_{k,min}}^{r_{k,max}} |g(r \cos \theta_k, r \sin \theta_k)|$$

where, $\phi_k, 1 \leq k \leq K_j$ represent the angular orientations of the quantization bins with respect to the VP $\overline{v}_j$. For robustness, bilinear interpolation is used to obtain the pixel gradients at sub-pixel locations, and the above computation of $D(k)$ is performed at sub-pixel resolution. In our experiments, $r$ varies in its range at half-pixel resolution. Examples of descriptors, obtained as above, by computing the edge strength in each angular bin, are shown in Fig. 3 for two different views of the same object.

### 2.4. Projective Transformation

Our motive behind constructing image descriptors is to perform matching of an object in images captured from different viewpoints. As each image is a 2D projection of the same real-world scene, there usually exists a geometrical relationship between the corresponding keypoints or edges in a pair of images. For example, there exists a homography relationship between images of planar facades of a building. Our observations suggest that there is an affine correspondence between the $D(k)$ values computed for 2 images of the same object. Below, we show that these observations have a theoretical justification. In particular, we show that the transformation of the angles between the image lines used in the binning step while building the descriptor, is approximately affine.

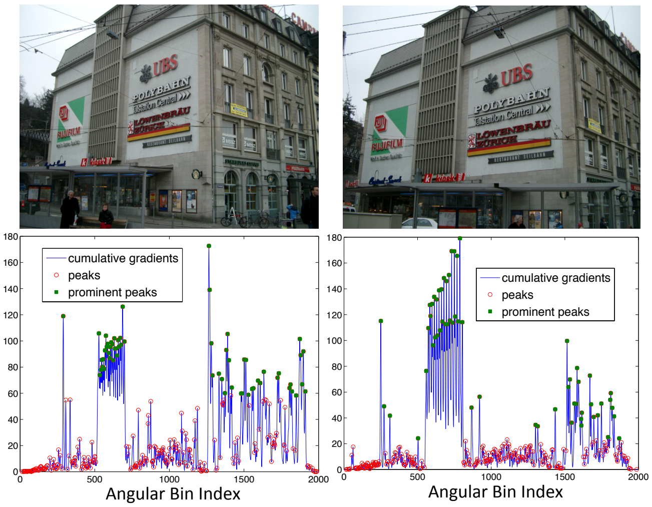

**Fig. 3**: Descriptors corresponding to a VP located above the image, for 2 views of a Manhattan scene. The prominent peaks aid in descriptor matching, as described in Section 3.

Consider two images (views) of the same scene consisting of a pencil of lines that pass through a vanishing point. Let the vanishing point for the first view be located at the origin. Using homogeneous representation, the $x$ and $y$ axes are given by $\mathbf{e}_x = (0\ 1\ 0)^T$ and $\mathbf{e}_y = (1\ 0\ 0)^T$. Using these vectors, any line $\mathbf{l}_\lambda$ is represented as

$$\mathbf{l}_\lambda = \mathbf{e}_x + \lambda \mathbf{e}_y = (\lambda\ 1\ 0)^T,$$

where $\lambda \in \mathbf{R}$. Without loss of generality, we assume that the inter-angle being studied is the angle between the $x$-axis and $\mathbf{l}_\lambda$. Observe that $\theta_\lambda = \tan^{-1}(-\lambda)$. Our goal is to show that the angle between the $x$-axis and $\mathbf{l}_\lambda$ undergoes an approximately affine transformation from one image to the other. To show this, denote the $3 \times 3$ homography between the two views using the matrix $\mathbf{H}$. In general, under the homography, the vanishing point is no longer at the origin for the second view, and $\mathbf{He}_x$ is no longer along the $x$-axis. Now, choose a transformation given by another $3 \times 3$ matrix $\mathbf{T}$ that translates the vanishing point back to the origin and rotates $\mathbf{He}_x$ back to the $x$-axis, as depicted in Fig. 4. Denote the $\mathbf{TH}$ transformation of $\mathbf{l}_\lambda$ by $\mathbf{l}_\gamma$, and the angle between $\mathbf{l}_\gamma$ and the $x$-axis by $\theta_\gamma$. Then,

$$\mathbf{l}_\gamma = \mathbf{THl}_\lambda = \mathbf{TH}(\lambda\ 1\ 0)^T = (a_1 + \lambda b_1\ a_2 + \lambda b_2\ 0)^T,$$

where, $\theta_\gamma = \tan^{-1} - \frac{a_1+\lambda b_1}{a_2+\lambda b_2}$ in which $(a_1, a_2, b_1, b_2)$ are the transformation parameters derived from the elements of $\mathbf{T}$ and $\mathbf{H}$. Under the assumption that the vanishing point is far away from the image, so that $\theta_{\max} - \theta_{\min}$ is small, we can use the Taylor series approximation $\tan^{-1}(\alpha) \approx \alpha$ where $\alpha$ is a small angle (expressed in radians).

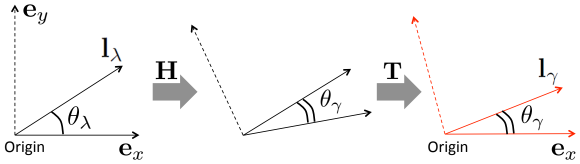

**Fig. 4**: Lines seen from one viewpoint are related to corresponding lines seen from a novel viewpoint via a projective transformation.

Accordingly,

$$\begin{aligned} \theta_\gamma &= -\frac{a_1 - \theta_\lambda b_1}{a_2 - \theta_\lambda b_2} \\ a_2\theta_\gamma &= -a_1 + b_1\theta_\lambda + b_2\theta_\gamma\theta_\lambda \end{aligned}$$

With the assumption of small inter-angles, the second order term $\theta_\gamma\theta_\lambda$ becomes negligibly small. If we neglect this cross term, then the transformation from $\theta_\lambda$ to $\theta_\gamma$ is approximately affine.

## 3. DESCRIPTOR MATCHING

An object in a Manhattan scene can have up to 3 VP's, and thus 3 descriptors. Hence, matching an object seen from two viewpoints without prior orientation information involves up to 9 pairwise matching operations. As seen above, the angular edge locations undergo an approximate affine transform with a change in viewpoint. It is necessary to invert this transformation before comparing the relative shapes of the edge strengths in the pair of descriptors being matched. A 2-step method is used to compare descriptors as described below.

### 3.1. Edge-wise corresponding mapping

To compute the approximate affine transform that morphs the descriptor between viewpoints, we exploit the fact that under the correct correspondence, pairs of coplanar edges generate approximately the same affine parameters, given by a scale-displacement pair $(s, d)$. Hence, a Hough transform-type voting procedure in the $(s, d)$ space for pairs of edges would result in a local maximum at the true scale $s^*$ and displacement $d^*$. Note that multiple local maxima will occur when the object has multiple planes supported by the VP directional axis. For robustness and efficiency, prominent edges are identified based on their edge strength. As shown in Fig. 3, edges with strength above a specified percentile threshold are chosen. Furthermore, for robustness to edge occlusion, only edges within close angular proximity are paired to cast votes, e.g., each prominent edge is paired with the $C$ closest edges.

Now, suppose that descriptor $D_1(k)$, $1 \leq k \leq K$ generates a set of $N_1$ peak pairs $(k_i, k_i')$, $1 \leq i \leq N_1$. Similarly, $D_2(m)$ generates a set of $N_2$ peak pairs $(m_j, m_j')$, $1 \leq j \leq N_2$. Now, pairs of peaks are cross-mapped between the two sets to generate votes for the $(s, d)$ histogram using $s = \frac{m_j' - m_j}{k_i' - k_i}$ and $d = m_j - sk_i$. To allow for angular inversion, i.e., top/bottom and left/right flipping around the VP, additional votes are generated by reversing the ordering of peaks within one of the above two sets. A coarse histogram of the $(s, d)$ votes can now be used to locate local maxima $(s^*, d^*)$, as shown in Fig. 5. The local maxima provide a relation between edges in the two views of the object. If a local maximum contains too few votes, a non-match is declared for that $(s^*, d^*)$ pair. If none of the local maxima contain enough votes, it is decided that the descriptors do not represent the same object.

### 3.2. Shape matching at corresponding edges

At each local maxima $(s^*, d^*)$, the local shape of the edge strength plot in the two descriptors being compared (e.g., the plots in Fig. 3) can be exploited to refine the matching process. Essentially, after compensating for the scaling factor $s^*$ and the displacement $d^*$, it remains to compare the shapes of the edge strength plots in the neighborhood of the edge pairs that voted for $(s^*, d^*)$.

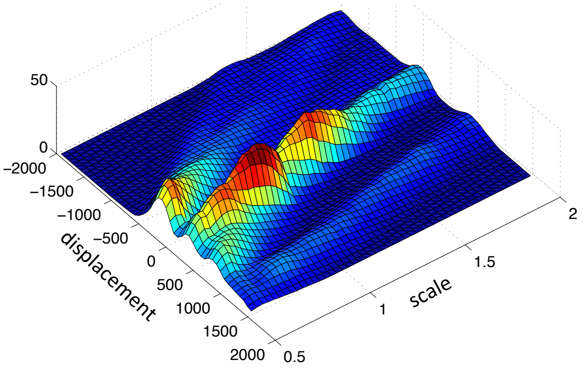

**Fig. 5**: A histogram is plotted to identify the scale and displacement at which two VP-based descriptors have the best match. This $(s, d)$ plot is generated using the 2 descriptors shown in Fig. 3.

To construct a metric for measuring the quality of the match, we perform the following steps for each prominent peak: (a) Consider the region in the angular neighborhood of the peak of the first descriptor (b) Compute the cumulative edge strength vector in this neighborhood, and normalize it such that the sum of all edge strengths is one. (c) Repeat this process for each matching prominent peak in the second descriptor, and (d) Compute for each pair of matching peaks —one taken from each descriptor— the absolute distance between the normalized cumulative edge strength vectors.

Finally, the absolute distances obtained in step (d) above are averaged across all matching peak pairs (possibly generated from multiple $(s, d)$ bins) and compared to a threshold. If the average distance between the normalized cumulative edge strength vectors is less than the threshold, a match is declared between the two descriptors.

## 4. EXPERIMENTAL RESULTS

The matching performance is evaluated on the ZuBuD database [15], which has 201 objects, each captured from 5 viewpoints. One of these views is chosen at random as the query, while the other four are included in the database. Cascaded 1D Hough transforms [12] are used for VP detection. For edge strength encoding, the angular range $[\theta_{\min}, \theta_{\max}]$ is quantized into 2000 uniform angular bins. Each VP descriptor occupies 8 kbits. With a 2.2 GHz Core i7 CPU running unoptimized MATLAB scripts, the 201 queries together required about 45 minutes for matching.

Before comparing two descriptors, prominent peaks are identified, for which the gradient strength is above the 80th percentile. See Fig. 3 for an example. For the purpose of voting into the scale-displacement histogram, peak pairs are obtained by considering each given prominent peak and 5 prominent peaks closest to it ($C = 5$). As each image has up to 3 VP descriptors, a match is declared if any one of the 3 descriptors of the query matches with any one of the 3 descriptors of the database image. The criterion for a match is the averaged absolute distance between the normalized cumulative edge strength vectors of the two descriptors as explained in Section 3.2.

A false positive is declared if a VP descriptor of a query object matches a descriptor from a different object in the database. The distribution of the pairwise average distances between the normalized cumulative edge strengths of query and database descriptors is shown in Fig. 6. By sweeping the match/non-match threshold, ROC curves are obtained as shown in Fig. 7, with an Equal Error Rate (EER) of 7%. When only the edge correspondences obtained from the local maxima of the scale-displacement plots (Section 3.1) are considered for matching, i.e., when the inter-descriptor distance is naively chosen as the absolute distance between the edge strengths compensated for scale and displacement, the matching performance is quite poor, with an EER of about 27.5%. This indicates that the shape matching step of Section 3.2 has significant discriminative ability. We limit our comparison to global scene descriptors of which GIST is a popular example. The GIST feature implementation [16] exploits aggregate image statistics but not scene geometry, resulting in inferior matching performance.

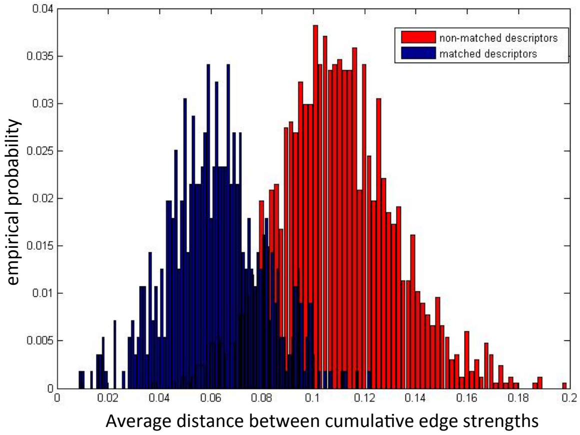

**Fig. 6**: Distribution of scores for true and false matches (ZuBud).

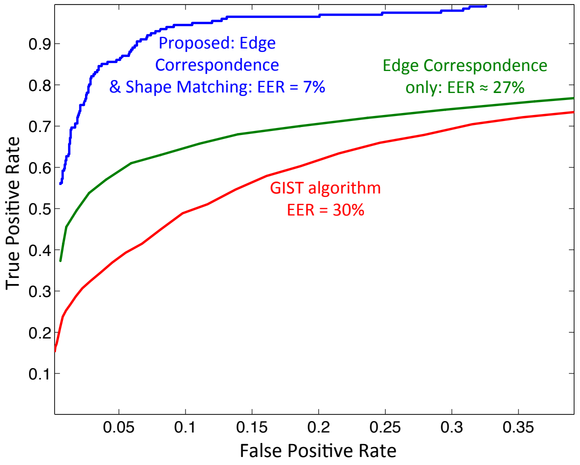

**Fig. 7**: ROC curves for descriptor matching algorithms.

## 5. SUMMARY

This work presents a global descriptor based on the locations and strengths of image edges in Manhattan scenes. The descriptor enables efficient storage and data transfer for querying. Unlike keypoint-based descriptors, the proposed descriptor holds perceptual relevance to the entire underlying object, i.e., the combination of the 3 VP descriptors intuitively yields a sketch of the underlying object. Shortcomings of this descriptor are in the computationally intense nature of the matching technique. Aside from reducing matching complexity, our ongoing focus is on using this descriptor for analyzing Manhattan scenes for computer vision applications.

## 6. REFERENCES

[1] D. Hubel and T. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, 1962.

[2] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346 – 359, June 2008.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, vol. 1, pp. 886–893.

[5] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, 2001.

[6] J. M. Morel and G. Yu, "ASIFT: A new framework for fully affine invariant image comparison," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 438–469, 2009.

[7] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, Oct 2003, pp. 1470–1477 vol.2.

[8] J. Coughlan and A. Yuille, "Manhattan world: Compass direction from a single image by bayesian inference," in *ICCV*, 1999.

[9] G. Schindler and F. Dellaert, "Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments," in *CVPR*, 2004.

[10] B. Caprile and V. Torre, "Using vanishing points for camera calibration," *International Journal of Computer Vision*, vol. 4, no. 2, pp. 127–139, 1990.

[11] J. P. Tardif, "Non-iterative approach for fast and accurate vanishing point detection," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1250–1257.

[12] B. Li, K. Peng, X. Ying, and H. Zha, "Vanishing point detection using cascaded 1d hough transform from single images," *Pattern Recognition Letters*, vol. 33, no. 1, pp. 1–8, 2012.

[13] M. Zuliani, C.S. Kenney, and B.S. Manjunath, "The multiRANSAC algorithm and its application to detect planar homographies," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, 2005, vol. 3, pp. III–153–6.

[14] J. Kosecka and W. Zhang, "Video compass," in *Computer Vision ECCV 2002*, Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, Eds., vol. 2353 of *Lecture Notes in Computer Science*, pp. 476–490. Springer Berlin Heidelberg, 2002.

[15] H. Shao, T. Svoboda, and L. V. Gool, "ZuBuD : Zurich Buildings database for image based recognition," Tech. Rep. 260, Computer Vision Lab, Swiss Federal Institute of Technology, Switzerland, Apr. 2003.

[16] A. Oliva and A. Torralba, "GIST Descriptor MATLAB Code," `http://people.csail.mit.edu/torralba/code/spatialenvelope/`.