Focusing Attention on Visual Features that Matter

Grace Tsai
gstsai@umich.edu
Benjamin Kuipers
kuipers@umich.edu

Electrical Engineering and Computer Science
University of Michigan
Ann Arbor MI 48109 USA

An indoor navigating agent needs to efficiently understand the geometric structure of its local environment in order to act. A common scene understanding approach is to generate a set of hypotheses about the geometric structure of the indoor environment and then test the hypotheses to select the one with the highest rank, from a single image [1, 4, 5, 6] or from a continuous stream of images (e.g. a video) [8, 9]. These methods simply detect features (e.g. lines [1, 4, 6], points [8, 9], and edges [6]) that are easily detectable for evaluating the hypotheses. In fact, some of the most informative features to discriminate the hypotheses may not be extracted if features are detected by fixed thresholds, since the informative regions may not have high image contrasts for features to be detected.

This paper demonstrates that by focusing attention on features in the informative regions, we can evaluate the hypotheses more efficiently. The idea of focusing on informative regions of the image space is inspired by the idea of saliency detection [2, 3, 7]. While these works typically define saliency regions based on image and motion properties of the pixels in the images [2, 3] or based on human fixations [7], our informative regions are defined in terms of the agent's own state of knowledge, the current set of hypotheses about the geometric structure of the indoor environment.

Given a set $\mathbf{M}$ of hypotheses, we divide the image into regions based on the expected information gain that each feature provides, which we call *informativeness* (Figure 1). We define the informativeness $I(p_j, \mathbf{M}) \in [0, 1]$ of point p_j, measuring its discriminating power among the set $\mathbf{M}$ as,

$$I(p_j, \mathbf{M}) = \log(|\mathbf{M}|) - H(\mathbf{M}^u | p_j), \tag{1}$$

where $H(\mathbf{M}^u | p_j)$ is the expected entropy of the set $\mathbf{M}$ with uniform prior.

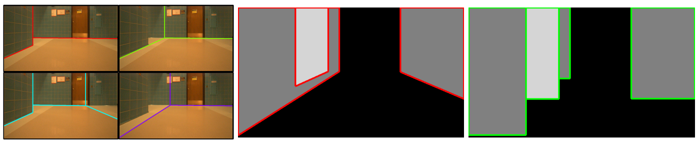

Figure 1: Example of the informative regions. (Best viewed in color.) (**Left**) The current set of hypotheses. (**Middle**) The gray-scale value reflects the informativeness $I(p_j, \mathbf{M}) \in [0, 1]$ of each pixel p_j in the current image based on the four hypotheses. Since the hypotheses are qualitatively distinctive, the image divides into several regions based on the informativeness. (**Right**) Since precisely computing the exact boundary of the informative regions can be computationally expensive, we use a set of axis-aligned boxes to approximate these regions. All points within each box are set to the same $I(p_j, \mathbf{M}) > 0$ value (maximum informativeness among all pixels within the box), and any point that is outside the boxes has $I(p_j, \mathbf{M}) = 0$.

To evaluate the set $\mathbf{M}$ of hypotheses, existing approaches extract high-contrast features across the entire image. However, efforts are being wasted when features with high image contrasts lie within uninformative regions, and opportunities may be missed when features in informative regions have relatively low image contrasts. Thus, in this paper, we adjust the threshold for extracting features in the informative regions to allow features to be extracted even if they have lower image contrasts. Moreover, when evaluating the hypotheses, instead of using all extracted features, we only use features that are capable of discriminating among the current set of hypotheses to reduce the computational cost.

We selected a Bayesian filter-based approach to scene understanding [8] to evaluate our attention focusing method. We compare the effectiveness of our method with the baseline [8] by computing the informativeness of the selected features at each frame. Figure 2 shows that with the same set of hypotheses, our method selects more features that are capable of discriminating the hypotheses and wastes no effort on features that are uninformative. In addition, our experimental results demonstrate that this bias of the search toward the most informative point features helps the Bayesian filter to converge to a single hypothesis more efficiently, without loss of accuracy. About 50% of the time, our method converges to a single hypothesis while only about 30% of the time, the baseline method converges to a single hypothesis.

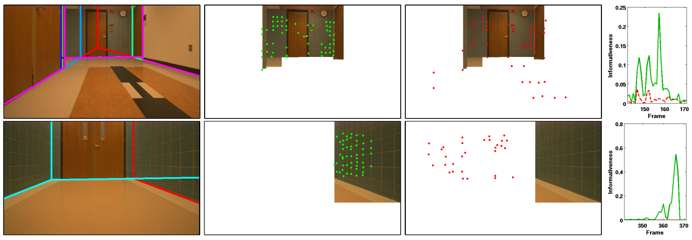

Figure 2: Examples of our attention focusing method. (Best viewed in color.) Each row is a snapshot from our experiment. The first column is the set of hypotheses $\mathbf{M}$ at that frame. The second column visualizes the informative point features (green) $\mathbf{P}_a$ that are selected by our method to evaluate the hypotheses. The third column shows the point features (red) from the baseline set $\mathbf{P}_b$, which are simply point features with high corner responses. For the second and the third column, only the informative regions are shown, and non-informative regions are shown in white. The last column is a comparison of the informativeness of using each feature set. Our proposed attention focusing method $I(\mathbf{P}_a, \mathbf{M})$ is shown in green solid lines, and the baseline method $I(\mathbf{P}_b, \mathbf{M})$ is shown in red dashed lines. Our method achieves higher informativeness because more point features that are capable of discriminating the hypotheses are tracked. In general, there are 1.5 to 6.5 times more point features that are capable of discriminating the hypotheses in the informative set $\mathbf{P}_a$ than in $\mathbf{P}_b$. In some extreme cases (second row), the baseline set $\mathbf{P}_b$ does not contain any features that are capable of discriminating the hypotheses so the informativeness $I(\mathbf{P}_b, \mathbf{M})$ at those frames are zero.

Our main contribution is to show that by using informativeness to control the process of feature acquisition, we can use computational resources more efficiently to discriminate among hypothesized interpretations of a visual scene, with no loss of accuracy. Informativeness allows our method to focus computational resources on regions in the scene where different hypotheses make different predictions. We demonstrate our method using the problem of real-time scene understanding for a mobile agent (e.g. [8, 9]), but it is equally applicable to other scene understanding problems (e.g. [1, 4, 5, 6]). Our experimental results demonstrate that this bias of search towards informative features provides more discriminating power among the hypotheses than simply using features that are easy to detect, with no loss of accuracy.

[1] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. *ICCV*, 2009.

[2] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. *CVPR*, 2007.

[3] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 1998.

[4] David Changsoo Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. *CVPR*, 2009.

[5] David Changsoo Lee, Abhinav Gupta, Martial Hebert, and Takeo Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. *NIPS*, 2010.

[6] Scott Satkin, Jason Lin, and Martial Hebert. Data-driven scene understanding from 3d models. *BMVC*, 2012.

[7] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 2009.

[8] Grace Tsai and Benjamin Kuipers. Dynamic visual understanding of the local environment for an indoor navigating robot. *IROS*, 2012. Dataset: `www.eecs.umich.edu/~gstsai/release/Umich_indoor_corridor_2012_dataset.html`.

[9] Grace Tsai, Changhai Xu, Jingen Liu, and Benjamin Kuipers. Real-time indoor scene understanding using Bayesian filtering with motion cues. *ICCV*, 2011.