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

Abstract

A common approach to scene understanding generates a set of structural hypotheses and evaluates these hypotheses using visual features that are easy to detect. However, these features may not necessarily be the most informative features to discriminate among the hypotheses. This paper demonstrates that by focusing attention on regions where the hypotheses differ in how they explain the visual features, we can then evaluate those hypotheses more efficiently. We define the *informativeness* of each feature based on the expected information gain that the feature provides to the current set of hypotheses, and demonstrate how these informative features can be selected efficiently. We evaluate our attention focusing method on a Bayesian filter-based approach to scene understanding. Our experimental results demonstrate that by focusing attention on the most informative point features, the Bayesian filter converges to a single hypothesis more efficiently, with no loss of accuracy.

1 Introduction

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, Lee et al. [5] generates hypotheses from image lines and evaluates the hypotheses based on the total number of pixels that each hypothesis agrees with an orientation map. Hedau et al. [2] generates scene layout candidates from image lines and ranks the candidates with a pre-trained predictor that scores the hypotheses based on global perspective cues. Lee et al. [6] searches through all possible room configurations (layout and box-like objects) and evaluates each configuration with a pre-trained scoring mechanism. Satkin et al. [7] uses a data-driven approach to generate layout hypotheses and learns a predictor to link image features with a 3D model library. For on-line mobile agent that perceives its local environment through a temporally continuous stream of images (e.g. a video), Tsai, et al. [11] generates a set of hypotheses from the first frame of the video, and uses a Bayesian filter to evaluate the hypotheses on-line based on their abilities to explain the 2D motions of a set of tracked features. Tsai and Kuipers [10] extended the real-time scene understanding method to generate children hypotheses on-line from existing hypotheses to describe the scene in more detail. These methods simply detect

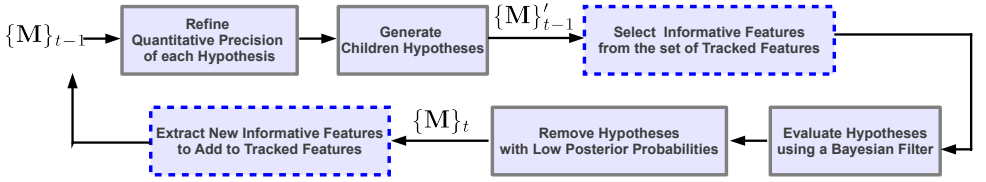


Figure 1: System pipeline. (Best viewed in color.) We demonstrate our attention focusing method in an on-line generate-and-test framework for scene understanding [10]. The steps with solid gray blocks are adapted from [10], and the steps with dashed blue blocks show where we select and extract the informative features. After generating children hypotheses ($|\{\mathbf{M}\}'_{t-1}| \geq |\{\mathbf{M}\}_{t-1}|$) , we select point features from the current set of tracked features that are informative to discriminate among the hypotheses $\{\mathbf{M}\}'_{t-1}$. Once the posterior probabilities of the hypotheses are updated and hypotheses with low probabilities are removed ($|\{\mathbf{M}\}_t| \leq |\{\mathbf{M}\}'_{t-1}|$), we identify new informative features based on the current set of hypotheses $\{\mathbf{M}\}_t$ to add into the tracking set. These features will be used to evaluate hypotheses in future frames.

features (e.g. lines [2, 5, 7], points [10, 11], and edges [7]) 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, because the most 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. We divide the image into regions based on the expected information gain that each feature provides, which we call *informativeness*. The idea of focusing on informative regions of the image space is inspired by the idea of saliency detection [3, 4, 8]. While these works typically define saliency regions based on image and motion properties of the pixels in the images [3, 4] or based on human fixations [8], 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. We adapt the threshold for extracting features for each region based on its informativeness. If a region is more informative, features with lower image contrasts are allowed to be used for hypotheses evaluation. We selected a Bayesian filter-based approach to scene understanding [10] to evaluate our attention focusing method. 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.

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. [10, 11]), but it is equally applicable to other scene understanding problems (e.g. [2, 5, 6, 7]).

2 Methodology

We demonstrate our attention focusing method on an on-line generate-and-test framework that constructs the geometric structure of the indoor environment [10]. Figure 1 describes our framework. We propose a method for selecting the set of point features $\mathbf{P}$ that are most informative for evaluating the set $\mathbf{M}$ of hypothesized models. The goal is to select features $\mathbf{P}$ that maximize the information gain $IG(\mathbf{M},\mathbf{P})$:

$$IG(\mathbf{M},\mathbf{P}) = H(\mathbf{M}) - H(\mathbf{M}|\mathbf{P}) \tag{1}$$

where $H(\mathbf{M})$ is the entropy of the current set of hypotheses and $H(\mathbf{M}|\mathbf{P})$ is the entropy given the set of point features $\mathbf{P}$. To explicitly maximize Equation 1, we need to evaluate the hypotheses with all combinations of all possible features, and then select the combination that returns a minimum expected entropy $H(\mathbf{M}|\mathbf{P})$. This process is very costly. However, we observe that a point feature will increase $IG(\mathbf{M})$ only if at least two hypotheses have different explanations about its 2D motion. In other words, a point p_j is "informative" if it lies in a region where at least two hypotheses make different predictions. We define $I(p_j,\mathbf{M}) \in [0,1]$ to be the *informativeness* of point p_j, measuring its discriminating power among the set $\mathbf{M}$,

$$I(p_j,\mathbf{M}) = \log(|\mathbf{M}|) - H(\mathbf{M}^u|p_j), \tag{2}$$

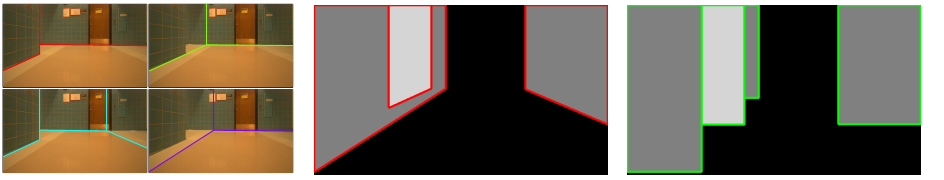
where $H(\mathbf{M}^u|p_j)$ is the expected entropy of the set $\mathbf{M}$ with uniform prior. Higher informativeness $I(p_j,\mathbf{M})$ means the point is able to provide larger information gain. If all hypotheses explain the 2D motion of point p_j in the same way, the point is not informative $I(p_j,\mathbf{M}) = 0$. Section 2.1 describes how the informativeness $I(p_j,\mathbf{M})$ of a point is computed, and Section 2.2 describes how we identify a set of informative points $\mathbf{P}$ for a set $\mathbf{M}$ of hypotheses..

2.1 Compute Informativeness of a Point Feature

For any point p_j in the image space, its informativeness $I(p_j,\mathbf{M})$ reflects how informative that point is for evaluating the current set of hypotheses $\mathbf{M} = \{M_1, M_2, ..., M_N\}$. $I(p_j,\mathbf{M}) \in [0,1]$ is positive if the point is capable of discriminating at least two hypotheses, and is zero if the point does not provide any information to discriminate among any hypotheses.

Given two hypotheses, if a point is informative $I(p_j,\mathbf{M}) > 0$, the two hypotheses have different explanations about its 2D motion. A hypothesis predicts the 2D motion of point p_j by reconstructing the point in 3D based on the 3D plane that the point is on, and then projects the point onto another frame [10, 11]. Thus, the key for the two hypotheses to have different predictions is when the two hypotheses assign the point to different 3D planes. If there is a difference from this pair of hypotheses, $I(p_j,\mathbf{M})$ of point p_j increases. At the end, $I(p_j,\mathbf{M})$ is the sum of scores from all possible pairs of the current hypotheses.

In fact, the informativeness $I(p_j,\mathbf{M})$ of all the points can be divided into several regions, where all points within each region have the same $I(p_j,\mathbf{M})$. Figure 2(b) is an example of these regions. For efficiency, instead of computing the precise boundaries of these regions, we approximate these regions with a set of non-overlapping boxes that specify which portions of the image are informative. The upper bounds of these boxes are the top image border. All points within each box are set to the same $I(p_j,\mathbf{M}) > 0$ value, and any point that is outside the boxes has $I(p_j,\mathbf{M}) = 0$. Figure 2(c) is an example of our box approximation. Note that it is possible that a point that originally has zero informativeness becomes non-zero in



(a) Hypotheses (b) Informative Regions (c) Box Approximation

Figure 2: An example of the informative regions and our box approximation for those regions. (Best viewed in color.) (a) The current set of hypotheses at this frame. (b) 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 shown in (a). Since the hypotheses are qualitatively distinctive, the image divides into several regions based on the informativeness. However, to precisely compute the exact boundary of these regions can be computationally expensive. Thus, we use a set of boxes to approximate these regions as shown in (c). All points within each box are set to the same $I(p_j, \mathbf{M}) > 0$ value, and any point that is outside the boxes has $I(p_j, \mathbf{M}) = 0$. The informativeness of each box is set to the maximum informativeness among all pixels within the box, so no information is lost by using the box approximation.

our box approximation, but all informative points remain informative. Thus, we do not lose any information by using this approximation.

Formally, we represent our box approximation based on a set of non-overlapping boxes $\{b_1, b_2, ..., b_{n_b}\}$. The informativeness $I(b_k, \mathbf{M})$ of each box b_k is proportional to the number of hypothesis pairs that the point can discriminate,

$$I(b_k, \mathbf{M}) = \begin{cases} \frac{1}{n_b(n_b-1)/2} \sum_{M_m, M_n \in \mathbf{M}} \delta(b_k, M_m, M_n) & \text{if } N > 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $\delta(b_k, M_m, M_n) \in \{0,1\}$ equals to 0 if hypotheses M_m and M_n are the same within box b_k, and equals to 1 if the hypotheses differ. Two hypotheses are the same if the associated 3D wall that is projected to the box area b_k is the same for the two hypotheses. We check whether the two walls are the same in 3D. Since the walls are perpendicular to the ground, a 3D wall is parameterized by a line $W = (\alpha, d)$ on the ground plane, where α is the orientation of the line which implies the normal direction of the wall plane in 3D, and d_j is the directed distance from the origin of the ground-plane map to the line. With this parameterization,

$$\delta(b_k, M_m, M_n) = \begin{cases} 0 & \text{if } |\alpha_m - \alpha_n| < \alpha_{same} \text{ and } |d_m - d_n| < d_{same} \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

where (α_m, d_m) and (α_n, d_n) are the walls in hypothesis m and n, respectively. α_{same} and d_{same} are the thresholds for considering the two walls to be the same. In our experiments, $\alpha_{same} = 0.00872$ radian and $d_{same} = 0.05$ meter.

To find the boxes, we start by finding their left and right bounds, and then find the lower bound. (The upper bound lies along the top image border.) The left and right bounds of the boxes correspond to a set of break points along the image columns. These break points only occur at the projected image locations of the vertical wall borders of the current hypotheses. We sort all the break points from the left to the right to form the bounds of the boxes, and form a set of candidate boxes using adjacent bounds. We then compute the informativeness

of each box using Equation 3, and remove boxes that have $I(b_k, \mathbf{M}) = 0$. For each informative box, the lower bound is the lowest horizontal line that encloses the ground-wall boundary segment of all hypotheses that pass through this box. Note, if the lowest horizontal line of a box is below the border of the image, the lower bound is set at the image border.

2.2 Select Informative Point Features

To evaluate the hypotheses $\mathbf{M} = \{M_1, M_2, ..., M_N\}$, [10, 11] extracts point features that have high corner responses from the entire image $\mathtt{I}_t$. The corner response $V(p_j)$ of a point p_j is defined as the minimum eigenvalue of the covariance matrix of derivatives over its neighborhood $S(p_j)$ [9]

$$\begin{bmatrix} \sum_{S(p_j)} (\frac{d\mathtt{I}_t}{dx})^2 & \sum_{S(p_j)} (\frac{d\mathtt{I}_t}{dx} \frac{d\mathtt{I}_t}{dy}) \\ \sum_{S(p_j)} (\frac{d\mathtt{I}_t}{dx} \frac{d\mathtt{I}_t}{dy}) & \sum_{S(p_j)} (\frac{d\mathtt{I}_t}{dy})^2 \end{bmatrix} \tag{5}$$

However, efforts are being wasted when points with high corner responses lie within uninformative regions, and opportunities may be missed when points in informative regions have relatively low corner responses. Thus, we need to adjust the threshold for extracting point features in the informative regions to allow point features to be extracted even if they have lower corner responses. Moreover, when evaluating the hypotheses, instead of using all the tracked features, we only use point features that are capable of discriminating among the current hypotheses to reduce the computational cost.

On the other hand, an informative point may not be a good feature to track due to low corner response. The tracking quality of the point will greatly affect the hypotheses evaluation process, because an ill-tracked point may not agree with the predicted 2D motion of a correct hypothesis. Thus, we introduce a cost term $C(p_j)$ to penalize the system for using point p_j,

$$C(p_j) = 1 - \frac{V(p_j)}{V_{max}} \tag{6}$$

where $V(p_j)$ is the corner response of point p_j, and V_{max} is the value of the maximum corner response from the current image $\mathtt{I}_t$.

Given a set of candidate point features $\mathbf{P}_c$ in the current image $\mathtt{I}_t$, we determine which points to be added into the tracking set for evaluating the hypotheses in a later frame. [1] (We will discuss how these candidate point features are extracted later.) Inspired by [1], the most efficient way to evaluate hypotheses is to use a diagnosis method that can well discriminate the hypotheses and has a low cost at the same time. Thus, to select the set of point features for evaluation, we maximize

$$\sum_{p_j \in \mathbf{P}_c} (I(p_j, \mathbf{M}) - C(p_j)) \delta(p_j) \tag{7}$$

where $\delta(p_j) \in \{0, 1\}$ equals to 1 if the point is selected to be tracked and 0 if the point is not going to be used. Maximizing Equation 7 is equivalent to selecting all the points that have more informativeness than cost $I(p_j, \mathbf{M}) > C(p_j)$. By maximizing Equation 7, a more informative point can be selected even with lower corner response, and a point that is less informative needs to have high corner response in order to be selected. For efficiency, in our experiments, we only allow at most 20 points to be added at each frame. If $\sum_{p_j \in \mathbf{P}_c} \delta(p_j) > 20$, we add 20 points with the highest gain $I(p_j, \mathbf{M}) - C(p_j)$.

[1] A point feature needs to be tracked for at least one frame in order to be used to evaluate the hypotheses.

The set of candidate point features $\mathbf{P}_c$ are extracted in the non-zero informativeness regions with a minimum corner response τ,

$$\tau = \min(V_{max}(1 - \max_{p_j}(I(p_j, \mathbf{M}))), \tau_{min}). \tag{8}$$

If the corner response is less than this threshold, it is impossible to be used based on Equation 7. We set a hard threshold τ_{min}, to avoid using unreliable points to ensure the quality of hypotheses evaluation. In our experiments, we set $\tau = 0.0000001$. In addition, a candidate point is not considered if that point is too close (less than 20 pixels) to an existing tracked point in the image space.

Besides the informative features, we also extract corner features with high corner responses as they become available because these features can potentially be informative for evaluating the hypotheses that are generated in future frames. For the same reason, we keep track of a point feature as long as it is trackable even when it is not informative for the current set of hypotheses. Thus, at each frame, we select the subset of the tracked points $\mathbf{P}_t$ with non-zero informativeness $I(p_j, \mathbf{M}) > 0$ to evaluate the hypotheses. To evaluate the hypotheses, we extract point correspondences between frame t_s and the current frame $t > t_s$. Given a hypothesis, we construct the 3D location of a point feature in the global frame of reference given its tracked location in t_s, and then, project the point onto the current frame t to compare with the observation, the tracked location of the point at frame t (re-projection error). The likelihood of that hypothesis is a function of the re-projection error. The likelihood function is more informative when t_s is larger, so we automatically adjust $t_s \in [5, 20]$ to ensure the number of features exceeds a threshold. [2]

3 Evaluations

We implemented our attention focusing method within the on-line generate-and-test framework of [10], and compared it with the baseline method [10], which simply uses point features with high corner responses. The evaluation was done using the Michigan Indoor Corridor 2012 Video Dataset [10]. Our implementation uses the same parameters for the two methods, except for those that are related to point feature extraction.

We compare the effectiveness of our method with the baseline [10] by computing the informativeness of the selected features at each frame. We define the informativeness $I(\mathbf{P}, \mathbf{M})$ of a set of point features $\mathbf{P}$ relative to a set of hypotheses $\mathbf{M}$ as

$$I(\mathbf{P}, \mathbf{M}) = \log(|\mathbf{M}|) - H(\mathbf{M}^u|\mathbf{P}). \tag{9}$$

where $\mathbf{M}^u$ is the set $\mathbf{M}$ of hypotheses at the current frame, but with uniform prior. We then compute the likelihood of each hypothesis based on $\mathbf{P}$, and update the posterior probabilities of the hypotheses. $H(\mathbf{M}^u|\mathbf{P})$ is the entropy of the posterior distribution.

We use the set of hypotheses $\mathbf{M}_a$ that exist at each frame when running the proposed method. At the meantime, we tracked two sets of point features. The first set $\mathbf{P}_b$ is obtained by simply tracking features with high corner responses as it is in [10], and the second set $\mathbf{P}_a$ is obtained by our proposed method where features are extracted in the informative regions even when their corner responses are low. [3] Point features in both sets may overlap. At

[2] We only use points that can be tracked for at least five frames to ensure that the point is reliable.

[3] Subscript $_a$ represents attention and $_b$ represents baseline.

Dataset	L	+	T 1	T 2	Overall
$I(\mathbf{P}_a,\mathbf{M}_a) > I(\mathbf{P}_b,\mathbf{M}_a)$	79.92%	63.41%	61.02%	72.99%	71.80%
$I(\mathbf{P}_a,\mathbf{M}_a) < I(\mathbf{P}_b,\mathbf{M}_a)$	10.81%	19.02%	20.34%	0%	10.76%
$\sum I(\mathbf{P}_a,\mathbf{M}_a)$	11.67	5.08	5.5	1.21	23.46
$\sum I(\mathbf{P}_b,\mathbf{M}_a)$	4.34	3.81	3.98	0.03	12.16
# Frames $\lvert\mathbf{M}_a\rvert > 0$	259	205	59	211	734
MAP $\mathbf{M}_a$ Accuracy	**94.31%**	93.8%	**92.8%**	**95.47%**	**94.12%**
Weighted $\mathbf{M}_a$ Accuracy	92.79%	**93.08%**	**92.57%**	**94.79%**	**93.35%**
$I(\mathbf{P}_a,\mathbf{M}_b) > I(\mathbf{P}_b,\mathbf{M}_b)$	70.62%	67.98%	52.16%	52.82%	61.71%
$I(\mathbf{P}_a,\mathbf{M}_b) < I(\mathbf{P}_b,\mathbf{M}_b)$	8.25%	17.98%	7.23%	2.82%	8.28%
$\sum I(\mathbf{P}_a,\mathbf{M}_b)$	34.04	13.88	6.06	21.36	75.34
$\sum I(\mathbf{P}_b,\mathbf{M}_b)$	11.48	4.45	4.9	0.9	21.73
# Frames $\lvert\mathbf{M}_b\rvert > 0$	303	228	69	390	990
MAP $\mathbf{M}_b$ Accuracy	94.30%	**94.03%**	91.65%	94.62%	93.68%
Weighted $\mathbf{M}_b$ Accuracy	**93.33%**	92.82%	92.27%	92.38%	92.67%
# Frames	341	311	391	411	1454

Table 1: Quantitative comparison with the baseline method [10]. The top half of the table is the results of running the proposed method, and the bottom half is the results of running the baseline. $I(\mathbf{P}_1,\mathbf{M}) > I(\mathbf{P}_2,\mathbf{M})$ reports the percentage of frames when feature set $\mathbf{P}_1$ is more informative than $\mathbf{P}_2$ among all the frames that have more than one hypothesis ($|\mathbf{M}| > 0$). $\sum I(\mathbf{P},\mathbf{M})$ reports the sum of informativeness over all the frames with $|\mathbf{M}| > 0$. MAP Accuracy is the average accuracy of the hypothesis with the highest posterior probability at each frame. Weighted Accuracy is the average weighted accuracy of the set of hypotheses at each frame, where the weight is the posterior probability of each hypothesis. (The dataset provides ground-truth labeling for every 10 frames.) See text for discussion.

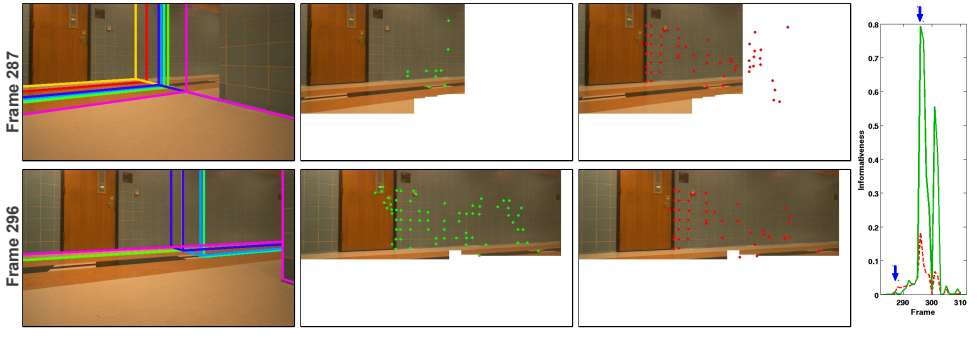


Figure 3: An example where the informative set $\mathbf{P}_a$ provides less informativeness than the baseline set $\mathbf{P}_b$. (Best viewed in color.) Since there is only one hypothesis prior to frame 287, our attention focusing method does not add in new points to the informative set $\mathbf{P}_a$ (green dots) prior to frame 287. At frame 287, a set of children hypotheses are generated so our method adds new informative features into the informative set $\mathbf{P}_a$. Once these points are tracked for some frames (second row), these informative points are used to discriminate the hypotheses. This phenomenon is reflected on the large increasing slope of informativeness $I(\mathbf{P}_a,\mathbf{M})$ (green solid line) between the two blue arrows. On the other hand, the baseline set $\mathbf{P}_b$ (red dots) continue to add in points with high corner responses so at frame 287 the baseline set happens to have more points that are capable of discriminating the hypotheses.

each frame, we compute the informativeness by using each set of features ($I(\mathbf{P}_b, \mathbf{M}_a)$ and $I(\mathbf{P}_a, \mathbf{M}_a)$) to evaluate the current set of hypotheses $\mathbf{M}_a$. We repeat the same comparison using the set of hypotheses $\mathbf{M}_b$ when running the baseline, and compute the informativeness ($I(\mathbf{P}_b, \mathbf{M}_b)$ and $I(\mathbf{P}_a, \mathbf{M}_b)$) at each frame. Notice that since $\mathbf{M}_a$ and $\mathbf{M}_b$ may consist of different hypotheses, the set of informative points $\mathbf{P}_a$ may be different in the two comparisons. However, the set of baseline features $\mathbf{P}_b$ in both runs are the same because these features are extracted independent of the hypotheses. These comparisons are shown in Table 1.

In most cases, the informative set $\mathbf{P}_a$ provides more informativeness than the baseline set $\mathbf{P}_b$. This is because more features in the informative set $\mathbf{P}_a$ lie in the region where the current hypotheses give different predictions than those in the baseline set $\mathbf{P}_b$. In some extreme cases, none of the baseline points lie in the informative regions. Figure 4 shows examples of these situations. Sometimes the baseline features $\mathbf{P}_b$ provide equal or more informativeness than the informative features $\mathbf{P}_a$. This happens at the first few frames when a large set of children hypotheses are generated. Because the informative set $\mathbf{P}_a$ is so focused on discriminating the parent hypotheses, $\mathbf{P}_a$ may not contain features that can discriminate the children hypotheses at the first few frames when they are evaluated. However, our method adds new informative features that discriminate the children hypotheses at the frame when they are generated, so after tracking these features for some times (at least 5 frames), $\mathbf{P}_a$ provides more discriminative power than $\mathbf{P}_b$. Figure 3 is an example of this situation.

Since the point features that are used in our methods and the baseline methods are different, the hypotheses that the two methods evaluated may differ. The total number of hypotheses evaluated in our method is larger than those in the baseline method. This is due to a threshold on the posterior probability for determining whether a hypothesis is good enough to generate children hypotheses. Our method evaluates the hypotheses more efficiently than the baseline and thus, more hypotheses exceeded this threshold and generated children hypotheses. Even though our method evaluates more hypotheses, our method converges to a single hypotheses more often. As shown in Table 1, 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.

In Table 1, we report the accuracy of our method and the baseline method based on our implementation. [4] The accuracy of our methods is similar to the baseline method. This suggests that by focusing attention on regions that are informative, regions where the current hypotheses have different explanations of the point features, we can converge to a single hypothesis more efficiently with no loss of accuracy.

4 Conclusion and Future Work

We demonstrate that by focusing attention on visual features that are informative, we can evaluate the hypothesized model of the scene more efficiently. A feature is informative if it is capable of discriminating among the hypotheses. In this paper, we define *informativeness* of a point feature mathematically and proposed method to identify informative features. We evaluate our attention focusing method on an on-line generate-and-test framework that constructs the geometric structure of the indoor scene [10]. 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

[4] Our implementation of the baseline method reaches similar accuracies as those reported in [10].

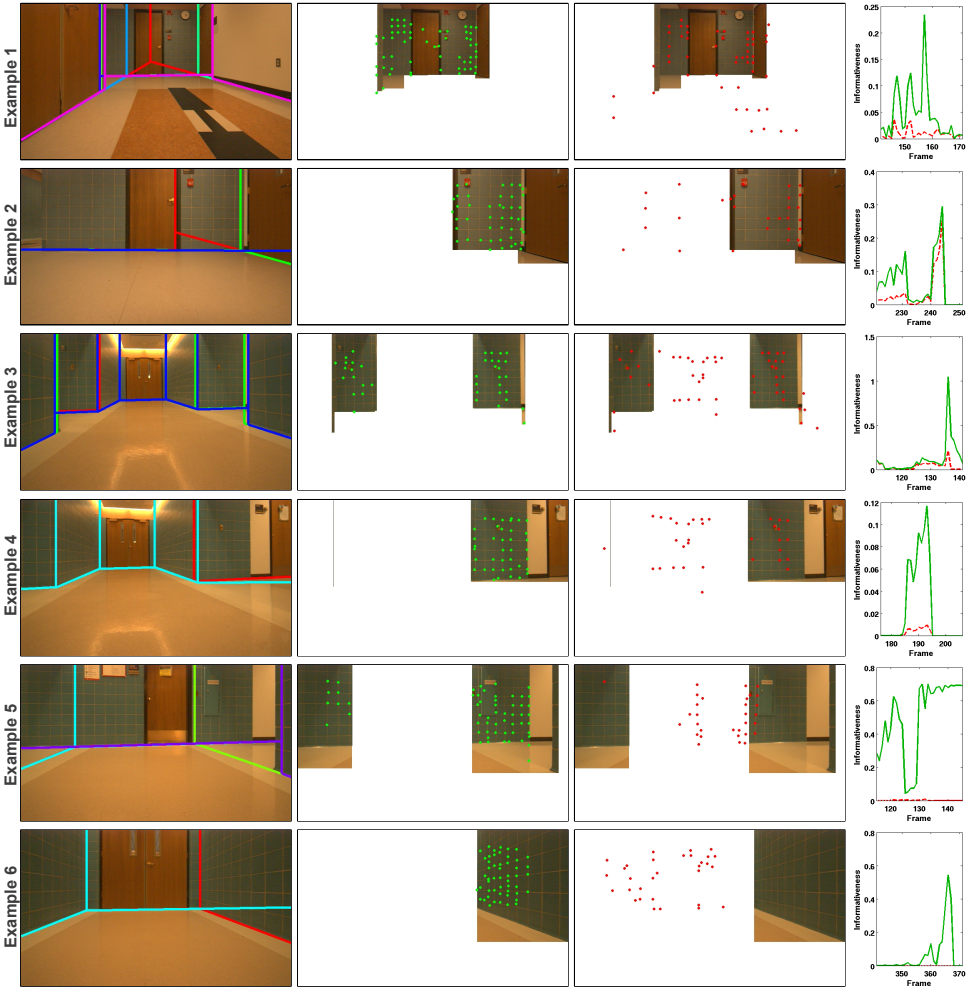


Figure 4: Examples of our attention focusing method. (Best viewed in color.) Each row is a snapshot of one of the four datasets. The first column is the set of hypotheses $\mathbf{M}$ at that frame. The second column visualizes the point features (green) that are used to evaluate the hypotheses from the informative set $\mathbf{P}_a$. The third column shows the point features (red) from the baseline set $\mathbf{P}_b$ that are visible at the current frame. For the second and the third column, only the informative regions ($I(b_k, \mathbf{M}) > 0$) are shown, and non-informative regions are shown in white. The last column is 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 (last row), the baseline set $\mathbf{P}_b$ does not contain any features to discriminate the hypotheses so the informativeness $I(\mathbf{P}_b, \mathbf{M})$ at those frames are zero.

no loss of accuracy. Our future work is to apply our attention focusing method to real-time active vision. We are designing a motion planner for the agent to increase the area of informative regions in the image space so that the agent can obtain more informative features to evaluate the hypotheses more efficiently.

Acknowledgment

This work has taken place in the Intelligent Robotics Lab in the Computer Science and Engineering Division of the University of Michigan. Research of the Intelligent Robotics lab is supported in part by grants from the National Science Foundation (CPS-0931474, IIS-1111494, and IIS-1252987).

References

[1] Johan De Kleer and Brian C Williams. Diagnosing multiple faults. *Artificial intelligence*, 32:97–130, 1987.

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

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

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

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

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

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

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

[9] Jianbo Shi and Carlo Tomasi. Good features to track. *CVPR*, 1994.

[10] 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.

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