Active Scene Recognition with Vision and Language

Xiaodong Yu*, Cornelia Fermüller†, Ching Lik Teo‡, Yezhou Yang‡, Yiannis Aloimonos‡
Computer Vision Lab, University of Maryland, College Park, MD 20742, USA
xdyu@umiacs.umd.edu*, fer@cfar.umd.edu†, {cteo, yzyang, yiannis}@cs.umd.edu‡

Abstract

*This paper presents a novel approach to utilizing high level knowledge for the problem of scene recognition in an active vision framework, which we call **active scene recognition**. In traditional approaches, high level knowledge is used in the post-processing to combine the outputs of the object detectors to achieve better classification performance. In contrast, the proposed approach employs high level knowledge actively by implementing an interaction between a reasoning module and a sensory module (Figure 1).*

Following this paradigm, we implemented an active scene recognizer and evaluated it with a dataset of 20 scenes and 100+ objects. We also extended it to the analysis of dynamic scenes for activity recognition with attributes. Experiments demonstrate the effectiveness of the active paradigm in introducing attention and additional constraints into the sensing process.

1. Introduction

The paradigm of Active Vision [1, 2, 3, 17, 19] had invigorated Computer Vision research in the early 1990s. The ideas were inspired by the observation that in nature vision is used by systems that are active and purposive. By studying visual perception in isolation, we often end up with more complicated formulations and under-constrained problems. Thus, the Active Vision paradigm proposed that visual perception should be studied as a dynamic and purposive process for active observers that can control their imaging mechanism. Most previous work in this paradigm was concerned with low level robot vision problems, and applied the ideas to shape reconstruction and navigational problems, such as motion estimation, obstacle avoidance, surveillance and path planning. Higher level tasks of scene understanding and recognition have not been sufficiently studied in this framework. These problems require combining high level knowledge and reasoning procedures with low-level image processing and a systematic mechanism for doing so.

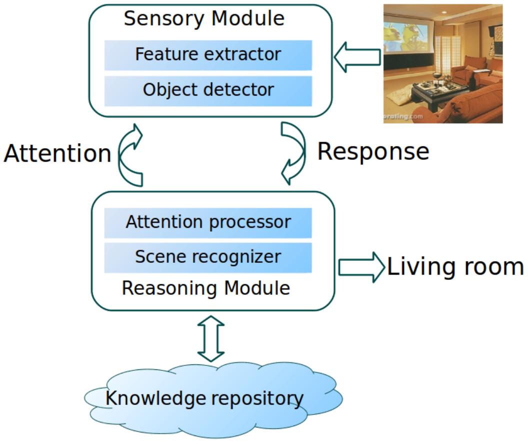


Figure 1. Overview of the active approach for scene recognition.

In this paper we propose a new approach to scene understanding in the paradigm of active vision. Central to the approach is a bio-inspired attention mechanism. Human perception is active and exploratory. We continuously shift our gaze to different locations in the scene. After recognizing objects, we will fixate again at a new location, and so on. Humans interpret visual input by using their knowledge of actions and objects, along with the language used for representing this information. It is clear that when we analyze a complex scene, visual processes continuously interact with our high-level knowledge, some of which is represented in the form of language. In some sense, perception and language are engaged in an interaction, as they exchange information that leads to meaning and understanding.

This idea is applied to the simplest interpretation problem, scene recognition, in this paper. The proposed system consists of two modules: (1) the reasoning module, which obtains higher level knowledge about scene and object relations, proposes attentional instructions to the sensory module and draws conclusions about the contents of the scene; (2) the sensory module, which includes a set of visual operators responsible for extracting features from images, detecting and localizing objects and actions. The novelty of the proposed active paradigm is that the sensory module does not passively process the image; instead, it is guided by the reasoning module, which decides what and where the sensory module should process next. Thus the sensory module shifts the focus of attention to a small number of objects at selected locations of the scene. This leads to faster and more accurate scene recognition.

Figure 1 illustrates the interaction between the two modules, which is modeled as an iterative process. Within each iteration, the reasoning module decides on what and where to detect next and expects the sensory module to reply with some results after applying the visual operators. The reasoning module thus provides a focus of attention for the sensory module, which can be an object to be detected and a place to be examined. For the problem of scene recognition, the interaction between the two modules is simple. (See Figure 6 for examples of the interaction over a given image.) However, our framework is more general. In Section 5 we discuss the extension of the framework to dynamic scene understanding. In this case the goal is to interpret the activity in the video. An activity is described by a set of quantities: the human, the tools, the objects, the motion, and the scene involved in the activity. Each of the quantities has many possible instances which can be described by their attributes (e.g., adjectives of nouns and adverbs of verbs). Thus the reasoning module at every iteration has to decide which quantity and which attribute to compute next. This procedure can be implemented in a hierarchical model of the proposed active scheme.

The rest of the paper is organized as follows: in the next section, we review related work; Section 3 describes an implementation of the proposed paradigm; in Section 4 we experimentally evaluate the active scene recognizer; in Section 5 we discuss how our framework can be generalized to object recognition and dynamic scene interpretation, and we demonstrate the ideas on the problem of recognizing hand activities on a small dataset; finally we draw conclusions in Section 6.

2. Related Works

Recognition by Components: The methodology for object, scene and activity recognition in this paper follows the idea of "recognition by components", which can be traced back to early work by Biederman [4]. In this methodology, scenes are recognized by detecting the inside objects [13], objects are recognized by detecting their parts or attributes [11], and activities are recognized by detecting the motions, objects and contexts involved in the activities [10]. However, all previous works employ passive approaches. As a result, they need to run through all object/attribute detectors over the testing images and videos before making the final conclusion. In this paper we explore an active approach, which aims at greatly reducing the number of object/attribute detectors needed for recognition of objects, scenes and activities.

Active Learning and Active Testing: Our work is a type of active testing and is closely related to the visual "20 question" game described in [5]. While the approach in [5] needs human annotators to answer the questions posed by the computer, our approach is fully automated without a human in the loop.

To select the optimal objects/attributes, we use the criterion of Maximum Information Gain, which have been widely used for active learning of objects and scenes [21, 25]. Information theory also have been used for object localization in application of face detection [22].

Employing Ontological Knowledge in Computer Vision System for Scene Interpretation: Ontological knowledge plays an important role in the reasoning and learning system of human. For example, in the problem of scene recognition, if we know that *coast* is a type of outdoor scene and also know that it is unlikely to find bookshelves therein. Hence, we do not need to apply the bookshelves detectors in the possible *coast* scene image. The work in [23] takes advantage of this type of knowledge in object detection. Similarly, the knowledge about objects and attributes is employed in [11]. Extending the knowledge about object hierarchy is employed in [15]. In this paper, we further explore the ontological knowledge about activities and attributes and present a pilot study using a hand activity dataset.

3. The Approach

3.1. System Overview

The proposed active scene recognizer classifies a scene by iteratively detecting the objects inside it. In the k-th iteration, the reasoning module provides an attentional instruction to the sensory module to search for an object O_k within a particular region L_k in the image. Then the sensory module runs the corresponding object detector and returns a response, which is the highest detection score d_k and the object's location l_k. The reasoning module receives this response, analyses it and starts a new iteration. This iteration continues until some terminating criteria are satisfied. To implement such an active scene recognizer, we need to implement the following components: (1)a sensory module for object detection; (2) a reasoning module for predicting the scene class based on the sensory module's responses; (3) a strategy for deciding which object and where in the scene the sensory module should process in the next iteration; and (4) a strategy for initializing and terminating the iteration. We will describe these components in the rest of this section.

3.2. Scene Recognition by Object Detection

In the proposed framework, the reasoning module decides the scene class S based on the responses X from the sensory module, which we call Scene Recognition by Object Detection (SROD). The optimal scene class of the given image belongs to the one that maximizes the probability:

$$S^* = \arg\max_{S \in [1:M]} p(S|X), \qquad (1)$$

where M is the number of scene classes.

The responses from the sensory module are a detection score and a detection bounding box. We only consider the objects' vertical positions, since they are more consistent within the images of the same scene class [23]. An object's vertical position is represented by a profile of the mask formed by the object's bounding box (see Figure 2 for an example). The object's mask formed by the object's bounding box is normalized to 256 × 256 pixels, and the profile is the histogram of pixels within the object's mask along the vertical axis. By this compact representation, we not only record the object's vertical location, but also record the object's scales along the horizontal and vertical axes. In the following, we denote this representation of an object's location as l_k.

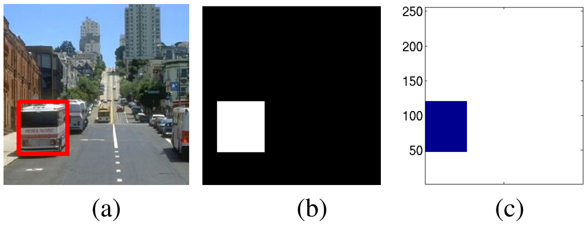
As described above, in each iteration, the sensory module returns a detection score d_i and a detected location l_i for the expected object O_i. Thus at step k, we have accumulated a list of detected score $d_{1:k}$ and corresponding locations $l_{1:k}$. Given $X = (d_{1:k}, l_{1:k})$, the probability of a scene S is :

$$\begin{aligned} P(S|X) &= p(S|d_{1:k}, l_{1:k}) \\ &\propto p(d_{1:k}, l_{1:k}|S) \\ &= p(d_{1:k}|S)p(l_{1:k}|S). \end{aligned} \qquad (2)$$

In the above equation, we assume $d_{1:k}$ and $l_{1:k}$ are independent given S. We approximate $p(d_{1:k}|S)$ by the inner product of $d_{1:k}$ and $\tilde{d}^S_{1:k}$, where $\tilde{d}^S_{1:k}$ is the mean $d_{1:k}$ of training examples for scene class S. Similarly, $p(l_{1:k}|S)$ is approximated by the inner product of $l_{1:k}$ and $\tilde{l}^S_{1:k}$. The advantage of this approximation is its simplicity and flexibility. We need to update the list of selected object in each iteration. If we adopt a parametric model for $p(d_{1:k}|S)$ and $p(l_{1:k}|S)$, we would need to learn the parameters for all permutations of $O_{1:k}, k = 1, ..., N$, where N is the total number of object categories in the dataset. For large N[1], such scheme would not work simply because of the computational constraints. Using a parameter-free approach, we avoid this difficulty.

3.3. Detecting Objects by The Sensory Module

The task of the sensory module is to detect the object required by the reasoning module and return a response. In this paper, we applied three object detectors: a Spatial Pyramid Matching object detector [12], a latent SVM object detector [8] and the texture classifier by Hoiem [9]. For each object class, we train all three object detectors and then select the one with the highest detection accuracy on a validation dataset to use in the test. Given a test image, the object detector will find a few candidates with corresponding detection scores. The one with the highest score is selected and sent to the reasoning module. The detection scores are normalized by Platt scaling [18] to obtain probabilistic estimates.



(a) (b) (c)

Figure 2. Representation of the object's location: (a) an object's bounding box; (b) the binary mask formed by the bounding box; (a) the profile of the object's binary mask along the vertical direction.

3.4. Attentional Instructions by The Reasoning Module

The interaction between the reasoning and sensory modules at iteration k starts from an attentional instruction issued by the reasoning module, based on its observation history. In this paper, the attentional instruction in iteration k includes *what* to look for, i.e., the object to detect (denoted as O_k) and *where* to look, i.e., the regions to detect (denoted as L_k). The criterion to select O_k and L_k is to maximize the expected information gain about the scene in the test image due to the response of this object detector:

$$\{O^*_k, L^*_k\} = \arg\max_{\substack{O_k \in \tilde{\mathcal{N}}_{k-1}, \\ L_k \in \mathcal{L}_k}} \mathrm{I}(S; d_k, l_k | d_{1:k-1}, l_{1:k-1}), \qquad (3)$$

where $\tilde{\mathcal{N}}_{k-1}$ denotes the set of indices of objects that have not been detected until iteration k, $\mathcal{L}_k$ denotes the search space of O_k's location. The global optimization procedure is approximated by two local optimization procedures. In the first step, we select O_k based on the maximum expected information gain criterion:

$$O^*_k = \arg\max_{O_k \in \tilde{\mathcal{N}}_{k-1}} \mathrm{I}(S; d_k, l_k | d_{1:k-1}, l_{1:k-1}). \qquad (4)$$

Then L^*_k is selected by thresholding $\mathbb{E}_S[\tilde{l}^S_{O^*_k}]$, the expected location of object O^*_k across all scene classes.

The expected information gain of O_k given the previous

[1] In our dataset, $N > 100$

response $d_{1:k-1}$ and $l_{1:k-1}$ is defined as:

$$\begin{aligned} &\mathrm{I}(S; d_k, l_k | d_{1:k-1}, l_{1:k-1}) \\ &= \sum_{d_k \in \mathcal{D}, l_k \in \mathcal{L}_k} p(d_k, l_k | d_{1:k-1}, l_{1:k-1}) \\ &\quad \times \mathrm{KL}[p(S|d_{1:k}, l_{1:k}), p(S|d_{1:k-1}, l_{1:k-1})]. \end{aligned} \tag{5}$$

The KL divergence on the right side of Equation 5 can easily be computed after applying Equation 2. To compute the first term on the right side of Equation 5, we factorize it as follows:

$$\begin{aligned} &p(d_k, l_k | d_{1:k-1}, l_{1:k-1}) \\ &\quad = p(d_k | d_{1:k-1}, l_{1:k-1}) p(l_k | d_{1:k}, l_{1:k-1}). \end{aligned} \tag{6}$$

The two terms on the right side of the above equation can be efficiently computed by their conditional probability with respect to S:

$$\begin{aligned} &p(d_k | d_{1:k-1}, l_{1:k-1}) \\ &= \sum_{S=1}^{M} p(d_k | S, d_{1:k-1}, l_{1:k-1}) p(S | d_{1:k-1}, l_{1:k-1}) \\ &= \sum_{S=1}^{M} p(d_k | S) p(S | d_{1:k-1}, l_{1:k-1}), \end{aligned} \tag{7}$$

where we assume d_k is independent of $d_{1:k-1}$ and $l_{1:k-1}$ given S. $p(d_k|S)$ can be computed by introducing the binary variable e_k, which indicates whether object O_k appears in the scene or not:

$$p(d_k|S) = \sum_{e_k \in \{0,1\}} p(d_k | e_k, S) p(e_k | S) \tag{8}$$

$$= \sum_{e_k \in \{0,1\}} p(d_k | e_k) p(e_k | S). \tag{9}$$

$p(e_k|S)$ encodes the high-level knowledge about the relationship between scene S and object O_k. One way to obtain it is to count the object labels in the training image set. Otherwise, we can obtain it from textual corpus. $p(d_k|e_k)$ encodes the information about the accuracy of different object detectors. The method to estimate its value is discussed in Section 4.1. The procedures described above are also employed to compute $p(l_k|d_{1:k}, l_{1:k-1})$ in a similar fashion.

Finally, we note that the expectation in Equation (5) needs to be computed at a set of sampling points of d_k (denoted as $\mathcal{D}$) and a set of sampling points of l_k (denoted as $\mathcal{L}_k$). $\mathcal{D}$ is within a one dimensional space between 0 and 1 and we draw samples of d_k uniformly. $\mathcal{L}_k$ can be parameterized by three parameters: the center position of O_k, y_k; the horizontal extent of O_k, w_k; and the vertical extent of O_k, h_k. We model these parameters by Gaussian distributions

$$y_k \sim \mathcal{N}(\mu_{y_k}, \sigma^2_{y_k}), \tag{10}$$

$$h_k \sim \mathcal{N}(\mu_{h_k}, \sigma^2_{h_k}), \tag{11}$$

$$w_k \sim \mathcal{N}(\mu_{w_k}, \sigma^2_{w_k}). \tag{12}$$

The means and variances of these Gaussian distributions are estimated from the training set. Thus the problem of drawing a sample of l_k becomes the problem of drawing a sample of y_k, h_k, w_k from three Gaussian distributions.

After drawing samples of d_k and l_k, we substitute them into Equation 5 to compute the expected information gain for O_k. Then among all possible O_k's, we select the object that yields the maximum expected information gain, O^*_k. The distribution of O^*_k's location in a particular scene S is approximated by $\tilde{l}^S_{O^*_k}$, which is computed as follows: first, we aggregate $l_{O^*_k}$ in all training samples of scene class S in the training stage; then we normalize the accumulated values into $[0, 1]$. Thus the expectation of $\tilde{l}^S_{O^*_k}$ across all scene classes, $\mathbb{E}_S[\tilde{l}^S_{O^*_k}]$, represents the distribution of O^*_k's location in an image of any scene class. Finally, we threshold this value by 0.5 and obtain a binary L^*_k, which provides the focus of attention for the sensory module in the next iteration.

3.5. Initializing and Terminating the Iteration

The interaction between two modules starts from the first object and its expected location, which are provided by the reasoning module. We select the object O_1 that maximizes the mutual information

$$O^*_1 = \arg\max_{O_1 \in [1:N]} \mathrm{I}(S; d_1, l_1). \tag{13}$$

To terminate the iteration, we can either stop after a fixed number of iterations (e.g., the 20 question game), or stop when the expected information gain at each iteration is below a threshold. In our experiments, we followed the former approach and found that 30 iterations are sufficient to produce competitive recognition results.

4. Experiments

4.1. Image Datasets

We evaluate the proposed approach using a subset of the SUN image set from [7]. Overall, the SUN dataset[7] contains 12K images, 1K scene classes and more than 200 object classes. We sort the scene classes by the number of examples and select the top 20 that have more than 50 examples per scene class. The remaining scene classes are discarded since they do not have sufficient number of examples to evaluate our algorithms. For each of the 20 selected scene classes, we randomly select 50 examples, where 30 of them are used for training and the rest 20 for testing. At the end, there are 127 object classes within our subset of SUN image set but only a handful of object classes appear in a particular scene class. As discussed in [7], a typical scene image contains seven object classes. The object detectors are trained using an additional dataset of 26,000 objects that is disjoint from the training/testing scene images as described in [7]. The obtained object detectors are then evaluated in the 600 scene training examples. The detection score, d_k, is normalized into $[0, 1]$ and evenly quantized into 10 discrete values. For each e_k (0 or 1), we accumulate the counts of d_k for each of its 10 values. Due to its discrete nature, $p(d_k|e_k)$ can be modeled as a multinomial distribution. Since Dirichlet is the conjugate prior for multinomial, we use a Dirichlet distribution $\text{Dir}(\alpha)$ as the prior for $p(d_k|e_k)$, where α represents the number of prior observations of d_k given a particular e_k. Through all experiments in this paper, we set the parameter $\alpha = 1$. We also tried other values of α and found no significant performance impact.

4.2. Performance of the Scene Recognizer

In the first experiment, we evaluate the scene recognizer (SROD) as described in Equation 2 while all objects are detected. The "ideal" SROD, where we use the object ground truths as the outputs of object detectors, is also evaluated to illustrate the upper limit of the performance of SROD. Three baseline algorithms are evaluated as listed below:

- SVM using GIST [16] features.
- SVM using Bag-of-Words (BoW). We used two types of local features, SIFT [14] and opponent SIFT [24], and the size of visual word dictionary is set as 500 for each of them.
- Classification and Regression Tree (CART) [6] that uses the object detection scores as attributes to predict the scene classes of a given image. The "ideal" CART, where the object ground truth is used as attributes, is also evaluated to illustrate the upper limit of the performance of CART.

Figure 3 compares the scene classification accuracy of these baseline algorithms and the SROD approach. The SROD approach significantly outperforms all the baseline algorithms. This result confirms the effectiveness of object-based approaches in interpreting complex scenes and the robustness of the SROD approach to the errors in object detection. It is worth to emphasize that there is still a lot of room to improve the current object-based scene recognizer, as suggested by the performance of the ideal SROD.

In addition, we evaluate the robustness of these scene recognition approaches with respect to the size of training samples. We randomly select a number of training examples from the training image set for each scene class and repeat the experiments three times. The mean and standard deviation of the average accuracy when using 5, 10, 15, 20, 25 and 30 training examples are reported in Figure 4. The proposed SROD method achieves substantially better performance than all baseline algorithms, including the CART algorithm that uses the same outputs of object detectors.

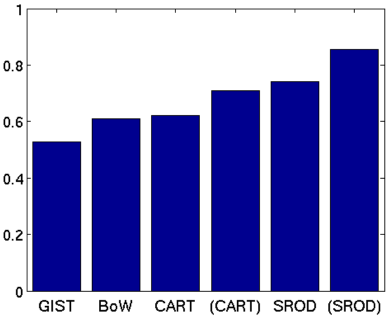


Figure 3. Comparison of scene classification accuracy of different approaches (GIST+SVM vs. BoW+SVM vs. CART vs. SROD). We also illustrate the "ideal" performance of CART and SROD, where we use the object ground truths as the outputs of object detectors. They are represented as "(CART)" and "(SROD)" in the figure respectively.

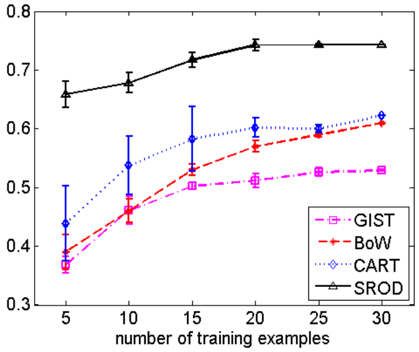


Figure 4. Classification accuracies of different approaches (GIST+SVM vs. BoW+SVM vs. CART vs. SROD) with respect to the number of training images.

4.3. Comparison of the Active Scene Recognizer vs. the Passive Scene Recognizer

In this experiment, we compare the proposed active scene recognizer with two baseline algorithms and the results are presented in Figure 5. Both baseline algorithms recognize scene class by iterative object detection, which is similar to the proposed SROD method. But they employ different strategies to select the to-be-detected object in each iteration. The first baseline (denoted as "DT" in Figure 5) follows a fixed object order, which is provided by the CART algorithm; while the second baseline (denoted as "Rand" in Figure 5) just randomly selects an object from the remaining object pool. Object selection obviously has a big impact on the performance of scene recognition, since

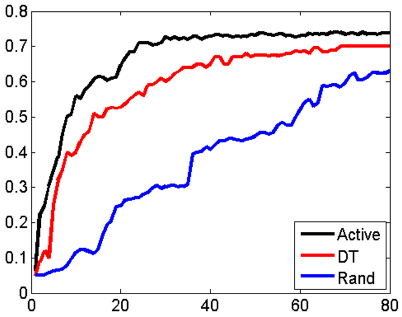


Figure 5. Comparison of classification accuracy among different object selection strategies (active vs. passive vs. random) in the object-based scene recognizers, with respect to the number of training images.

both the proposed active approach and the "DT" approach significantly outperform the "Rand" approach. The result also shows that the active approach is superior to the "DT" approach that is passive: the active approach can achieve stable performance after selecting 30 objects while the passive "DT" approach needs 60 objects. Furthermore, the object's expected location provided by the reasoning module in the active approach not only reduces the spatial search space to be about 1/3 to 1/2 of the whole image but also reduces the false positives in the sensory module's response. As a result, the proposed active approach achieves 3% to 4% performance gain compared to the passive approaches.

4.4. Visualization of the Interaction between the Sensory Module and the Reasoning Module

Figure 6 illustrates a few iterations of the active scene recognizer performed on a test image. It shows that after detecting twenty objects, the reasoning module is able to decide the correct scene class with high confidence.

5. Dynamic Scene Recognition

There are two key premises in the proposed active scheme: (1) a quantity can be recognized by accumulating evidences from its components; (2) the components can be assumed to be independent given the quantity. Given these two premises, the active scheme can be applied to select a small number of components to recognize the quantity without impairing the performance. In the previous section, we have applied this active scheme to recognize static scenes. However, this active scheme can also be applied to recognize objects by their parts and recognize activities by their motion and object properties.

In this section, we will demonstrate the application of the active scheme in an activity recognition problem. A big challenge in this problem is that the components are heterogeneous. While static scenes only involve a single quantity, i.e., objects, activities are described by different quantities, including motion, objects and tools, scenes, temporal properties, etc. To alleviate this problem, we propose a hierarchical active scheme for dynamic scene recognition. Figure 7 presents this method. In this scheme, each iteration performs the following four steps: (1) using the maximum information gain criterion, the activity-level reasoning module sends an attentional instruction to the quantity-level reasoning module that indicates the desired quantity (e.g., motion or objects); (2) the quantity-level reasoning module then sends an attentional instruction to the sensory module that indicates the desired attributes (e.g., object color/texture, motion properties); (3) the sensory module applies the corresponding detectors and returns the detector's response to the quantity-level reasoning module; (4) finally, the quantity-level reasoning module returns the likelihood of the desired quantity to the activity-level reasoning module.

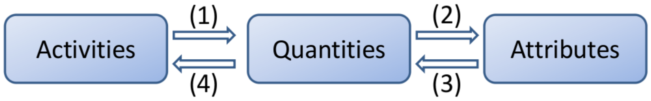


Figure 7. Hierarchical active scheme for dynamic scene recognition, where each iteration invokes four steps. Section 5 discusses the details.

Quantity	Attribute	$e = 1$	$e = 0$
Tools	Color	silver	other colors
	Texture	bristle	non-bristle
	Elongation	yes	no
	Convexity	yes	no
Motion	Frequency	high	low
	Motion variation	large	small
	Motion spectrum	sparse	non-sparse
	Duration	long	short

Table 1. Activity attributes in the hand activity dataset.

To demonstrate this idea, we used 30 short video sequences of 5 hand actions from a dataset collected from the commercially available PBS *Sprouts* craft show for kids (the hand activity data set). The activities are *coloring*, *drawing*, *cutting*, *painting*, and *gluing*. 20 sequences were used for training and the rest for testing. Two quantities are considered in recognizing an activity: the characteristics of tools and the characteristics of motion. Four attributes are defined for the characteristics of tools, including *color*, *texture*, *elongation*, and *convexity*; and four attributes are defined for the characteristics of motion, including *frequency*, *motion variation*, *motion spectrum*, and *duration*. The details of these quantities and attributes are described in Table 1.

The sensory module includes detectors for the 8 at-

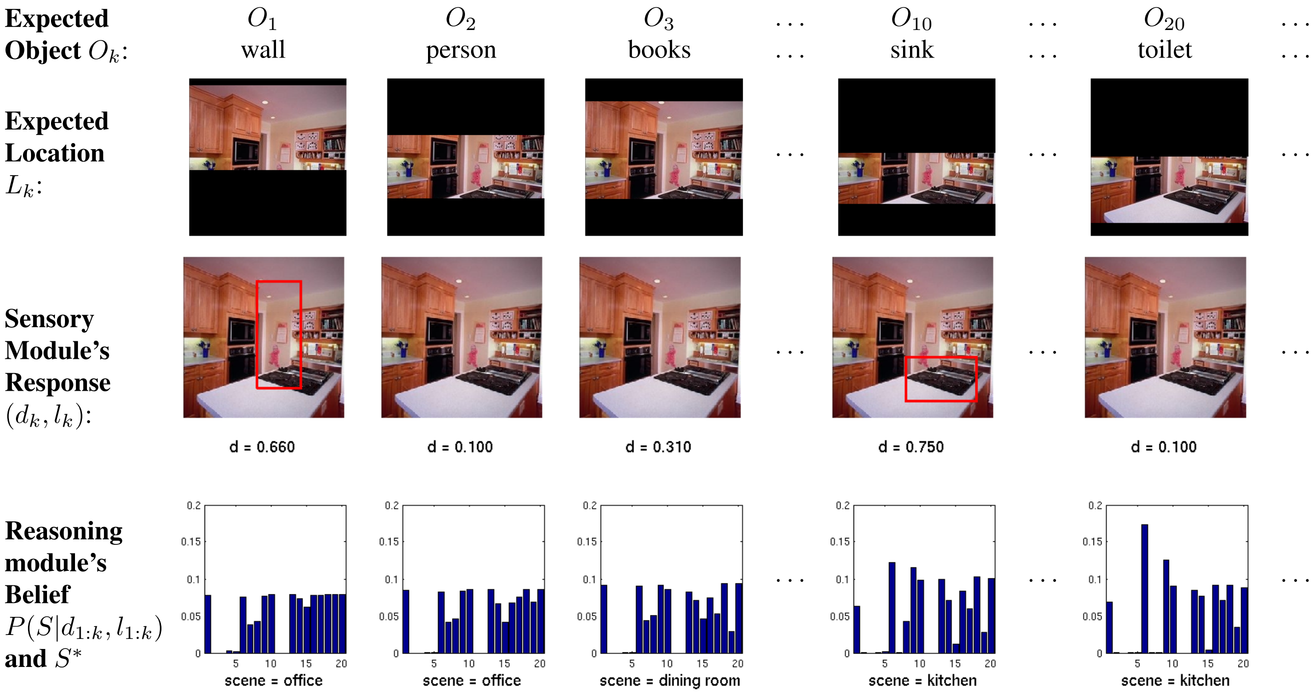


Figure 6. Visualization of the iterations between the reasoning module and the sensory module in an active scene recognition process. The detected regions with detection score greater than 0.5 are highlighted with a red bounding box.

tributes of tools/motion. To detect these attributes, we need to segment the hand and tools from the videos. Figure 8 illustrates these procedures, which are described as follows:

1. Hand regions S_h are segmented by applying a variant of the color segmentation approach based on Conditional Random Fields (CRF) [20] using a trained skin color model. Similarly, moving regions of hands and tools, S_f, are segmented by applying another CRF over the optical flow fields.

2. A binary XOR operation is applied on S_h and S_f to remove the moving hand regions and produce a segmentation of tools, S_T.

3. Apply a threshold t_f to remove regions with flows that are different from the hand regions and obtain a candidate region for tool, $\hat{S}_r$.

4. Detect edges in $\hat{S}_r$.

5. Fitting a minimum volume ellipse over the edge map of $\hat{S}_r$, which estimates the region of the detected tool.

Figure 9 shows the estimated ellipse enclosing the detected tool over some sample image frames from the dataset. This ellipse, together with $\hat{S}_r$, is then used as a mask to compute object-related attributes. The color and texture attributes were computed from histograms of color and wavelet-filter outputs, and the shape attributes were derived from region properties of the convex hull of the object and the fitted ellipse. The motion attributes were computed from the spectrum of the average optical flow over the sequence and the variation of the flow.

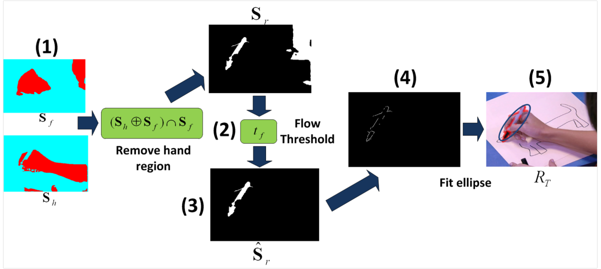


Figure 8. Procedures to extract hands and tools from the hand activity video sequence. Please refer to the text for details.

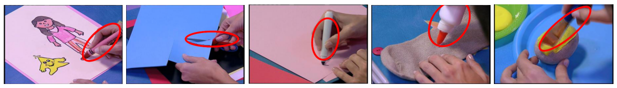


Figure 9. Sample frames for 10 testing videos in the hand action dataset: (from left to right) coloring, cutting, drawing, gluing, painting. The detected tool is fit with an ellipse.

Table 2 shows the interactions between the reasoning modules and the sensory modules for one of the testing videos. Here the sensory module only needed to detect two attributes before the reasoning module arrived at the correct conclusion. Overall, 8 out of 10 testing videos were recog-

Iteration	1	2	3	4
Expected quantity	Tools	Tools	Tools	Motion
Expected attribute	Elongation	Color	Texture	Duration
Sensory module's response	0.770	1.000	0.656	0.813
Reasoning module's conclusion	Coloring	Painting	Painting	Painting
Reasoning module's confidence	0.257	0.770	0.865	0.838

Table 2. An example of interactions between the reasoning module and the sensory module for hand activity recognition, where the ground truth of the activity class is *painting*.

nized correctly after detecting two to three attributes, while the remaining two testing videos could not be recognized correctly even after detecting all the attributes. This is because of errors in the segmentation, the choice of attributes and the small set of training samples.

6. Conclusion and Future Work

We proposed a new framework for scene recognition within the active vision paradigm. In our framework, the sensory module is guided by attentional instructions from the reasoning module and employs detectors of a small set of objects within selected regions. The attention mechanism is realized using an information theoretic approach, with the idea that every detected object should maximize the added information for scene recognition. Our framework is evaluated in a static scene dataset and shows the advantage over the passive approach. Also we discussed how it can be generalized to object recognition and dynamic scene analysis, and gave a proof of concept by implementing it for attribute based activity recognition.

In the current implementation, we have assumed that objects are independent given the scene class. Though this assumption simplifies the formulation, this is not necessarily true in general. In the future, we plan to remove this assumption and design a scene recognition model that better represents the complex scenes in the real world. Also, we will perform a comprehensive study of the proposed approach using larger image/video datasets to investigate the impact of the active paradigm.

References

[1] J. Aloimonos, I. Weiss, and A. Bandopadhay. Active Vision. *IJCV*, 2:333–356, 1988. 1

[2] R. Bajcsy. Active Perception. *Proceedings of the IEEE*, 76:996–1005, 1988. 1

[3] D. H. Ballard. Animate Vision. *Artificial Intelligence*, 48:57–86, 1991. 1

[4] I. Biederman. Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94:115–147, 1987. 2

[5] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual Recognition with Humans in the Loop. In *ECCV*, 2010. 2

[6] L. Breiman, J. Friedman, C. J. Stone, and R. Olshen. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984. 5

[7] M. J. Choi, J. Lim, A. Torralba, and A. S. Willsky. Exploiting Hierarchical Context on a Large Database of Object Categories. In *CVPR*, 2010. 4, 5

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. *PAMI*, 32(9):1627 – 1645, 2010. 3

[9] D. Hoiem, A. Efros, and M. Hebert. Automatic Photo Pop-up. In *ACM SIGGRAPH*, 2005. 3

[10] N. Ikizler-Cinbis and S. Sclaroff. Object, Scene and Actions: Combining Multiple Features for Human Action Recognition. In *ECCV*, 2010. 2

[11] H. N. Lampert, C. H. and S. Harmeling. Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer. In *CVPR*, 2009. 2

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, 2006. 3

[13] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification. In *NIPS*, 2010. 2

[14] D. G. Lowe. Distinctive Image Features from Scale-invariant Keypoints. *IJCV*, 20:91–110, 2004. 5

[15] M. Marszalek and C. Schmid. Semantic Hierarchies for Visual Object Recognition. In *CVPR*, 2007. 2

[16] A. Oliva and A. Torralba. Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope. *IJCV*, 42:145–175, 2001. 5

[17] J. olof Eklundh, P. Nordlund, and T. Uhlin. Issues in Active Vision: Attention and Cue Integration/Selection. In *BMVC*, pages 1–12, 1996. 1

[18] J. C. Platt. Probabilities for SV Machines. In *Advances in Large Margin Classifiers*, 1999. 3

[19] R. D. Rimey and C. M. Brown. Control of Selective Perception Using Bayes Nets and Decision Theory. *IJCV*, 12:173–207, 1994. 1

[20] C. Rother, V. Kolmogorov, and A. Blake. GrabCut: interactive Foreground Extraction using Iterated Graph Cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004. 7

[21] B. Siddiquie and A. Gupta. Beyond Active Noun Tagging: Modeling Contextual Interactions for Multi-Class Active Learning. In *CVPR*, 2010. 2

[22] R. Sznitman and B. Jedynak. Active Testing for Face Detection and Localization. *PAMI*, 2010. 2

[23] A. Torralba. Contextual Priming for Object Detection. *IJCV*, 53(2):153–167, 2003. 2, 3

[24] van de Sande, K. E. A., T. Gevers, and C. G. M. Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *PAMI*, 32(9):1582–1596, 2010. 5

[25] S. Vijayanarasimhan and K. Grauman. Cost-Sensitive Active Visual Category Learning. *IJCV*, 2010. 2