Visual Scene Interpretation as a Dialogue between Vision and Language

Xiaodong Yu and **Cornelia Fermüller** and **Yiannis Aloimonos**
University of Maryland Institute for Advanced Computer Studies, College Park, MD 20742-3275
xdyu@umd.edu, {fer, yiannis}@umiacs.umd.edu

Abstract

We present a framework for semantic visual scene interpretation in a system with vision and language. In this framework the system consists of two modules, a language module and a vision module that communicate with each other in a form of a dialogue to actively interpret the scene. The language module is responsible for obtaining domain knowledge from linguistic resources and reasoning on the basis of this knowledge and the visual input. It iteratively creates questions that amount to an attention mechanism for the vision module which in turn shifts its focus to selected parts of the scene and applies selective segmentation and feature extraction. As a formalism for optimizing this dialogue we use information theory. We demonstrate the framework on the problem of recognizing a static scene from its objects and show preliminary results for the problem of human activity recognition from video. Experiments demonstrate the effectiveness of the active paradigm in introducing attention and additional constraints into the sensing process.

Introduction

There has been a recent interest in research on scene and video understanding with a number of efforts devoted to introducing additional higher-level knowledge about image relationships into the interpretation process (Lampert and Harmeling 2009; Marszalek and Schmid 2007; Galleguillos and Belongie 2010). Current studies usually get this additional information from captions or accompanying text. It has been realized, however, that language in principle can be used to obtain additional high level information. Linguists and computational linguists have a longstanding interest in modeling lexical semantics, i.e. conceptual meanings of lexical items and how these lexical items relate to each other (Cruse 1986) and have created resources where information about different concepts, such as cause-effect, performs-functions, used-for, and motivated-by, can be obtained. For example, the WordNet database relates words through synonymy (words having the same meaning, like ”argue” and ”contend”) and hypernymy (”is-a” relationships, as between ”car” and ”vehicle”), among many others (Miller and Fellbaum 2007). Linguistics also has created large text corpuses and statistical tools so we can obtain probability distributions for the co-occurrence of any two words, such as how likely a certain noun co-occurs with a certain verb.

Using these linguistic tools, how we can aid vision to build better systems for interpreting images and video? As is well known computational vision is very challenging, and especially the tools available for solving recognition tasks are very limited. One way to use linguistic information is as a contextual system that provides additional information to the interpretation. For example, certain objects are likely to co-occur, such as ”tables” often co-occur with ”silverware” and ”glasses. We then can apply visual classifiers for individual objects to the images and use the output of these classifiers together with the context information in some minimization functions to produce the interpretation result. We call this the passive approach. Another way is to use linguistic information in an active reasoning system. Let’s say we are having a kitchen scene. Because we have prior knowledge about kitchens, their structure and the actions taking place in them and a large part of this knowledge is expressed in language, we can utilize this information during visual inspection. A knife in the kitchen will most probably be used for ”cutting” a food item, so vision can look for it. Of course there are many more relations that language can provide for prediction. We can search for the ”red big” <noun>, or describe the characteristics of the <noun> between <noun1> and <noun2>.

This paper describes a new framework for implementing this interaction between vision and language that draws its inspiration from human vision. Central to the approach is a bio-inspired attention mechanism. Human perception is active and exploratory. We actively shift our attention and give semantic significance to visual input on the fly by using our knowledge of images, actions and objects, along with the language we use for structuring our knowledge. In some sense, perception and language are engaged in a dialogue, as they exchange information that leads to meaning and understanding. In a similar way, we propose to implement a computational system that produces a semantic description of static and dynamic scenes.

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 con-

tents 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. 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 objects and actions to be detected, attributes to evaluated, and a place to be examined.

What are the advantages of this active paradigm? First it gives efficiency and accuracy. By providing prompt feedback from the language/reasoning module, the vision module can greatly reduce its search space so that it can focus on a small set of selected visual processes (classifiers, segmentation procedures) over selected regions within an image. Thus the image can be processed faster and more accurate. Second, the reasoning module can obtain organized higher-level information about object and action attributes and their relations (from adjectives, adverbs and preposition in language) and this information can be used to facilitate the vision processes by guiding the attention and introducing additional constraints for the segmentation and recognition. For example, it is easier to segment the long red object than to generally perform segmentation of the scene. Thirdly, since the reasoning module can automatically obtain high-level knowledge from language resources, the proposed approach can recognize scenes that it has never seen before.

In this paper the framework has been applied to two problems. First, we implemented the simplest interpretation problem, static scene recognition. A scene is described by the objects in it, and the reasoning module has to decide in every iteration on what object and where to look for in the scene. Second, we demonstrate preliminary results on the problem of dynamic scene understanding, where the goal is to interpret the activity in a video. An activity is described by a set of quantities, such as 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 this paper is organized as follows: In the next section we review related work. Then we describe the scene recognition system and evaluate it experimentally. Next we discuss the generalization of the framework to dynamic scene interpretation and a first implementation. Finally we draw conclusions and discuss future work.

Related Works

Recognition by Components: The methodology for object, scene and action recognition in this paper follows the idea of "recognition by components", which can be traced back to early work by Biederman (Biederman 1987). In this methodology, scenes are recognized by detecting the their objects (Li et al. 2010), objects are recognized by detecting their parts or attributes (Lampert and Harmeling 2009), and actions are recognized by detecting the motions, objects and contexts involved in the actions. However, all previous works employ passive approaches, while ours is active.

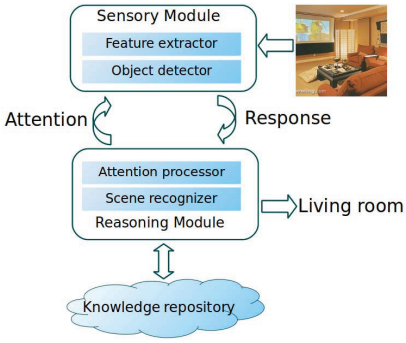


Figure 1: Overview of the active approach for scene recognition.

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 (Branson et al. 2010). While the approach in (Branson et al. 2010) 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 has been widely used for active learning of objects and scenes (Siddiquie and Gupta 2010; Vijayanarasimhan and Grauman. 2010).

Ontological Knowledge in Computer Vision System for Scene Interpretation: (Torralba 2003) uses knowledge about image features across the scene in object detection. Similarly, (Lampert and Harmeling 2009) exploits knowledge about object and attributes. (Marszalek and Schmid 2007) use knowledge about semantic hierarchy for object recognition. In this paper, we further explore the ontological knowledge about action and attributes in a pilot study of a hand action dataset.

The Approach

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 of the image L_k. 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 provide 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 and terminating the iteration. We describe these components in the rest of this section.

Scene Recognition by Object Detection

In the proposed framework, the reasoning module decides on the scene class based on the responses from the sensory module. The responses are a detection score and a location given by a detection bounding box. We only consider the objects' vertical positions, which are more consistent within the images of the same scene class (Torralba 2003).

At step k, we have a list of detected score $d_{1:k}$ and corresponding object locations $l_{1:k}$. Given these information, the probability of a scene S is :

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

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 of training examples of 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 optimal scene class of the given image is to the one that maximizes the probability:

$$S^* = \arg\max_{S\in[1:M]} p(S|d_{1:k}, l_{1:k}). \quad (2)$$

The Sensory Module

We applied three object detectors: a Spatial Pyramid Matching object detector (Lazebnik, Schmid, and Ponce 2006), a latent SVM object detector (Felzenszwalb et al. 2010) and a texture classifier (Hoiem, Efros, and Hebert 2005). For each object class, we train all three object detectors and choose the one with the highest detection accuracy

Attentional Instructions by The Reasoning Module

The task of the reasoning module is to provide an attentional instruction to the sensory module based on the observation history, $d_{1:k-1}$ and $l_{1:k-1}$. 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 gained about the scene in the test image due to the response of this object detector:

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

where $\tilde{\mathcal{N}}_{k-1}$ denotes the set of indices of objects that have not been asked at time 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:

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

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

The expected information gain of O_k given the response of previous detections $d_{1:k-1}$ and $l_{1:k-1}$ is defined as:

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

Next we describe in detail how to compute (5). The KL divergence can be computed from equation (1). To compute the first term in the right side of Equation (5), we factorize it as:

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

The two terms at the right hand side can be efficiently computed from their conditional probability with respect to S

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

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 event variable e_k, which indicates whether object O_k appears in the scene or not:

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

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

$p(e_k|S)$ encodes the high-level knowledge about the relationship between scene S and object O_k. We can obtain using statistics on textual corpus. $p(d_k|e_k)$ is computed from the training set as a posterior of a multinomial distribution with a Dirichlet prior, and $p(l_k|d_{1:k}, l_{1:k-1})$ can computed in a similar way.

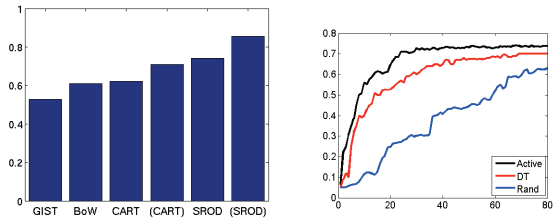
Finally, we note that the expectation in Equation (5) needs to be computed at a set of sampling points of d_k and a set of sampling points of l_k After drawing samples of d_k and l_k, we substitute them into Equation (5) to compute the information gain for O_k. Then among all possible O_k's, we select the object that yields the maximum information gain, O^*_k. Finally, L^*_k is selected by thresholding $\mathbb{E}_S[\tilde{l}^S_{O^*_k}]$.

Initializing and Terminating the Iteration

The first object chosen is the one that maximizes the mutual information

$$O^*_1 = \mathrm{argmax}_{O_1\in[1:N]}\mathrm{I}(S; d_1, l_1). \quad (10)$$

To terminate the dialogue, we can either stop after asking a fixed number of questions (e.g., the 20 question game), or stop when the information gain at each iteration is below a threshold.



(a) (b)

Figure 2: (a) Comparison of classification performance of different approaches. (b) Classification performance w.r.t number of object detectors.

Experiments

Image Datasets

We evaluated the proposed approach using a subset of the SUN images from (Choi et al. 2010). There is a total of 20 scenes and 127 objects in our dataset. For each scene, we select 30 images for training and 20 images for testing. The object detectors are trained using a dataset that is separated from the training/testing scene as described in (Choi et al. 2010).

Performance of the Scene Recognizer

In the first experiment, we evaluated the scene recognizer (SROD) as described in Equation 1 against the "ideal" SROD, which uses the objects' ground truths as the outputs of the detectors, and the following three methods:

- SVM using GIST features (Oliva and Torralba 2001)
- SVM using Bag-of-Words (BoW) with SIFT (Lowe 2004) and opponent SIFT (van de Sande, Gevers, and Snoek 2010) as local features.
- Classification and Regression Tree (CART) (Breiman et al. 1984) using the object detection scores as features. In addition we evaluated the "ideal" CART, where the object ground truth is used as features, to illustrate the uppper limit of CART's performance.

The performance of the different methods is shown in Figure 2a. Both object-based approaches, i.e., CART and SROD, outperform the approaches using holistic features, i.e, GIST and BoW. This result confirms the effectiveness of object-based approaches in interpreting high-level visual tasks such as scene recognition. It is worth emphasizing 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 a second experiment we compared the object selection strategy of the proposed active scene recognizer with two other methods as shown in Figure 2b. Both comparison algorithms use the same SROD formulation but employ different strategies to select the object in each iteration. The first method (denoted as "DT") follows a fixed object order, which is provided by the CART algorithm, and the second method (denoted as "Rand") randomly selects an object from the remaining object pool. As can be seen from the graph, object selection obviously has a big impact in the performance of scene recognition. Both, the proposed active approach and the "DT" approach significantly outperform the "Rand" approach, and the active approach is superior to the passive "DT" approach: the active approach can achieve competitive 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 1/3 to 1/2 of the image, but also reduces the false positives in the sensory module's response and yields a 3% to 4% performance gain compared to the passive approach.

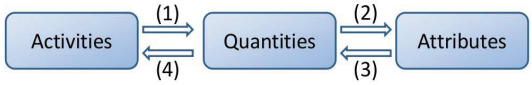


Figure 4: Hierarchical active scheme for dynamic scene recognition, where each iteration invokes four steps: (1) attentional instruction from the activity-level reasoning module; (2) attentional instruction from the quantity-level reasoning module; (3) responses from attribute detectors; (4) responses from the quantity-level reasoning module.

Visualization of the Dialogue between the sensory module and the reasoning module

Figure 3 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.

A Demo of an Active Video Parser

In language we can describe a manipulation activity by a number of quantities, such as the humans, the tools, the object and the action involved. The beauty is that these symbols (i.e. the quantities) that we have in language space to describe the action have direct correlates in visual space. That is, we can extract humans, tools, objects, motion patterns using vision. In the current implementation we only consider tools and actions. We describe them visually by first segmenting the corresponding image regions, i.e. the hands and the tool in the video sequence, and then we characterize them by attributes.

A big challenge for our formalism on the problem of activity recognition is that the components are heterogeneous. While static scenes only involve a single quantity (the objects), activities are described by different quantities (here the tools and actions). To alleviate this problem, we propose a hierarchical active scheme for dynamic scene recognition. Figure 4 presents this method. In this scheme, each iteration invokes 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 detectors response to the the quantity-

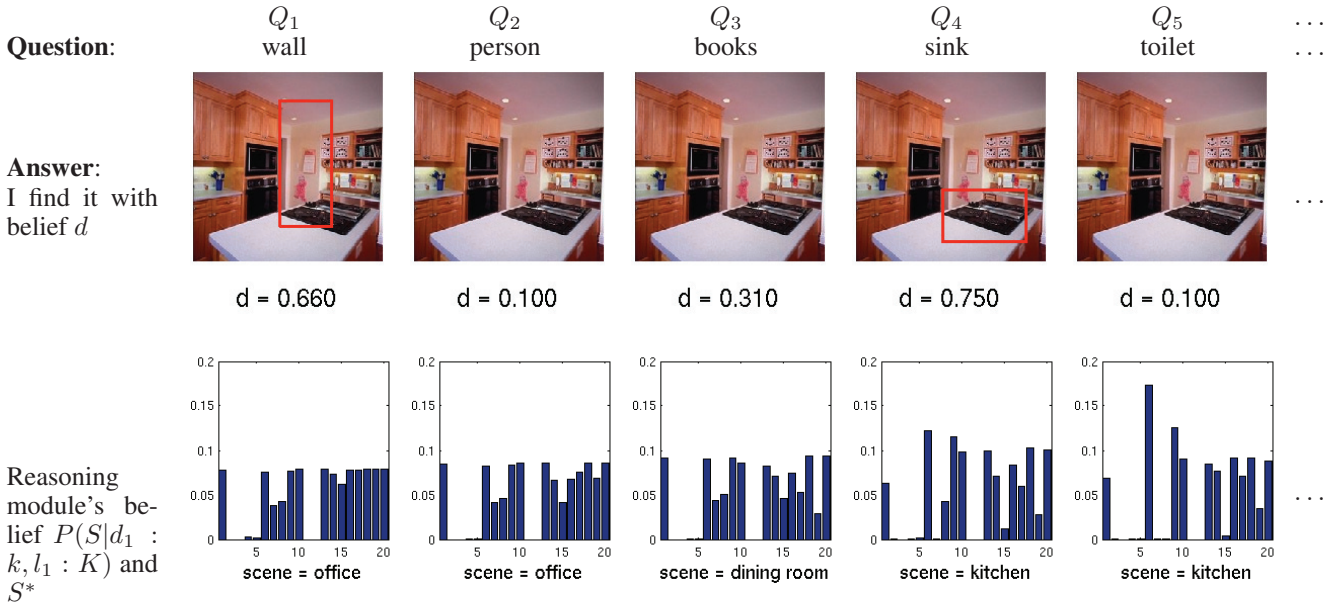


Figure 3: Visualization of the cognitive dialogue between the reasoning module and the sensory module for scene recognition, which starts from the object with the highest information gain, *wall*. The detected regions of objects with detection score greater than 0.5 are highlighted with a red bounding box.

level reasoning module; (4) finally, the quantity-level reasoning module returns the likelihood of the desired quantity to the activity-level reasoning module.

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 action data set). The actions 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 the 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 sensory module includes detectors for the 8 attributes of tools/motion. But before we detect the attributes, we need to segment the hand and tools from the videos. Figure 5 illustrates and explains these procedures.

Figure 6 shows the estimated ellipse enclosing the detected tool over some sample image frames from the dataset. This ellipse is then used as a mask to detect 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 adverbs of the actions were computed from the spectrum of the average estimated flow over the sequence and the variation of the flow.

Table 1 shows the dialogue for one of the testing videos. Here the reasoning module only required 2 questions before arriving at the correct conclusion. Overall, 8 out of 10 testing videos were recognized correctly after asking 2 to

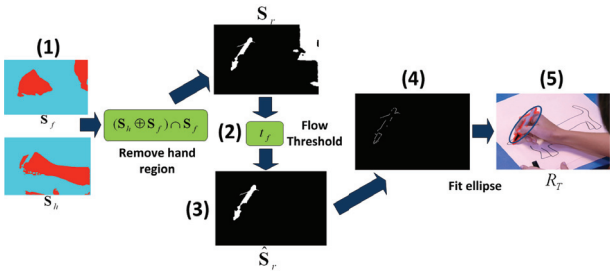


Figure 5: Procedures to extract hands and tools from the hand action video sequence. (1) First hand regions and moving regions are segmented using a CRF approach (**?**) based in the former case on color and in the latter on image flow. (2) Then applying a threshold to remove regions with flow that are different from the hand region, a region containing the tool is obtained (3). (4) Then edge detection is performed and (5) the best ellipse over the edge fragments is fitted to locate the tool.

3 questions, while the remaining 2 testing videos could not be recognized correctly even after asking all the questions. This is because of errors in the segmentation, the choice of attributes and the small set of training samples.

Conclusion and Future Work

We proposed a new framework for scene recognition combining language and vision. Current approaches to scene interpretation are passive and use language simply as a contextual system. In our active paradigm vision and language engage in a dialogue. Vision is guided by attention from language and sequentially carries out specific processes that are

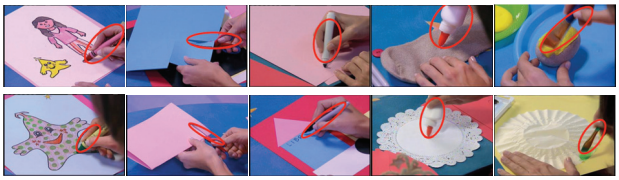


Figure 6: Sample frames for 10 testing videos in the hand action dataset. Frames in the same column belong to the same action class: (from left to right) coloring, cutting, drawing, gluing, painting. The detected tool is fit with an ellipse.

Iteration	1	2	3	4
Expected quantity	Tool	Tool	Motion	Motion
Expected attribute	Elongation	Color	Texture	Duration
Sensory modules 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 1: An example of the interaction between the reasoning module and the sensory module for hand action recognition, where the ground truth of the action class is *painting*.

initiated by language. The control of the dialogue is realized using an information theoretic approach, with the idea that every visual process should maximize the added information for scene recognition. We implemented and tested our framework for static scene recognition, and gave a proof of concept by implementing it for attribute based action recognition.

In future applications we will extend the proposed approach to activity recognition to additional quantities, such as the manipulated objects, the scene and transformations of the object during action. We will also study how to obtain attributes from language that we can map to vision space

Acknowledgments:

The support of the European Union under the Cognitive Systems program (project POETICON) and the National Science Foundation under the Cyberphysical Systems Program, (grant CNS-1033542) is gratefully acknowledged.

References

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

Branson, S.; Wah, C.; Babenko, B.; Schroff, F.; Welinder, P.; Perona, P.; and Belongie, S. 2010. Visual recognition with humans in the loop. In *ECCV*.

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

Choi, M. J.; Lim, J.; Torralba, A.; and Willsky, A. S. 2010. Exploiting hierarchical context on a large database of object categories. In *CVPR*.

Cruse, D. A. 1986. *Lexical semantics*. Cambridge, England: University Press.

Felzenszwalb, P.; Girshick, R.; McAllester, D.; and Ramanan, D. 2010. Object detection with discriminatively trained part based models. *PAMI* 32(9):1627 – 1645.

Galleguillos, C., and Belongie, S. 2010. Context based object categorization: A critical survey. *CVIU*.

Hoiem, D.; Efros, A.; and Hebert, M. 2005. Automatic photo pop-up. In *ACM SIGGRAPH*.

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

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, 2169–2178. Washington, DC, USA: IEEE Computer Society.

Li, L.-J.; Su, H.; Xing, E. P.; and Fei-Fei, L. 2010. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*.

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

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

Miller, G. A., and Fellbaum, C. 2007. *WordNet then and now*, volume 41. Springer. 209–214.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV* 42:145–175.

Siddiquie, B., and Gupta, A. 2010. Beyond active noun tagging: Modeling contextual interactions for multi-class active learning. In *CVPR*.

Torralba, A. 2003. Contextual priming for object detection. *IJCV* 53(2):153–167.

van de Sande, K. E. A.; Gevers, T.; and Snoek, C. G. M. 2010. Evaluating color descriptors for object and scene recognition. *PAMI* 32(9):1582–1596.

Vijayanarasimhan, S., and Grauman., K. 2010. Cost-sensitive active visual category learning. *IJCV*.