FOCUS: A Generalized Method for Object Discovery for Robots that Observe and Interact with Humans

Manuela M. Veloso, Paul E. Rybski, and Felix von Hundelshausen
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA, USA
mmv@cs.cmu.edu, prybski@cs.cmu.edu, felix@cs.cmu.edu

ABSTRACT

The essence of the *signal-to-symbol problem* consists of associating a symbolic description of an object (e.g., a chair) to a signal (e.g., an image) that captures the real object. Robots that interact with humans in natural environments must be able to solve this problem correctly and robustly. However, the problem of providing complete object models *a priori* to a robot so that it can understand its environment from any viewpoint is extremely difficult to solve. Additionally, many objects have different uses which in turn can cause ambiguities when a robot attempts to reason about the activities of a human and their interactions with those objects. In this paper, we build upon the fact that robots that co-exist with humans should have the ability of observing humans using the different objects and learn the corresponding object definitions. We contribute an object recognition algorithm, FOCUS, that is robust to the variations of signals, combines *structure* and *function* of an object, and generalizes to multiple similar objects. FOCUS, which stands for *Finding Object Classification through Use and Structure*, combines an activity recognizer capable of capturing how an object is used with a traditional visual structure processor. FOCUS learns *structural properties* (visual features) of objects by knowing first the object's *affordance properties* and observing humans interacting with that object with known activities. The strength of the method relies on the fact that we can define multiple aspects of an object model, i.e., structure and use, that are individually robust but insufficient to define the object, but can do when combined.

Categories and Subject Descriptors: I.2.10 Vision and Scene Understanding : Perceptual Reasoning

General Terms: Algorithms

Keywords: Functional object recognition, learning by demonstration

1. INTRODUCTION

"One can only see what one knows" is a key idea that emphasizes the importance of prior knowledge for sensor-based object recognition. The object models given to an intelligent robot dictate how it can take a stream of data, such as images from a camera, and extract meaningful information. One way to obtain new information about an environment is to learn through observation. Learning through observation is a powerful technique by which a robot can obtain knowledge about the physical world by watching humans interact with objects within it. Specifically, such observations provide a powerful method for learning affordance properties [12] of those objects. Affordance properties, or how an object can be interfaced with, capture the essence of an object's utility. For instance, a table can be used as a place to place objects, but it can also be used as something to stand on. Likewise, a step ladder can be used to reach high objects but objects can also be placed on it. In this work, we add information related to the *use* or *function* of the object with the aim of recognizing and generalizing objects robustly in any environment.

In traditional object recognition approaches, models of objects to be observed are given to an intelligent sensor as a mapping of features of the sensory signal to object descriptions. For example, CAD models and size and edge descriptors may be used to define and recognize objects. The sensory data stream must be searched for features of these objects given a number of assumptions about the kinds of objects, their placement, and the position of the sensor with respect to those objects. The signal-to-symbol problem is extremely challenging not only due to the difficulty of defining a general model of a specific object (e.g., a model of a specific kind of object) but also due to the brittleness of the mapping to a general signal (e.g., any image conditions).

As an example of the complexity of this problem, Figure 1 shows a visual representation of a set of different kinds of chairs. Note the difference in three-dimensional structure and volume, as well as how the orientation of the chair with respect to the observer can change its visually observed image. One approach to recognizing these different chairs out of an image might involve storing three-dimensional models of exemplar chairs (such as CAD models) and inferring the correct orientation of the chair as well as the relative position of the observing camera to the chair. This complexity increases as the number of different kinds of possible chairs increases. Further complexity is added by considering that the range to the chair is unknown, that there may be other objects occluding the image of the chair, and the fact that the image of the chair has to be extracted from a noisy and cluttered background image. In addition, there will be many objects of interest besides chairs. As we now well know, the complexity and size of the prior models can quickly spiral out of control. However to intelligently process sensory signals, it is critical to reason about objects and therefore the absolute need to invest on reliable object recognition.

Our algorithm, FOCUS (*Finding Object Classification through Use and Structure*), models inanimate objects in the environments by structural and functional definitions. The structural part of the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
HRI'06, March 2–4, 2006, Salt Lake City, Utah, USA.
Copyright 2006 ACM 1-59593-294-1/06/0003 ...$5.00.

Figure 1: An example of the complexity of modeling the visual notion of "chair". Chairs come in many different shapes and sizes and their appearance can change drastically as the viewpoint of the camera changes.

model aims at capturing a simple and generalized definition of an object (e.g., a chair has some color). The functional part of the model captures how one uses an object: *one sits down on a chair*. FOCUS includes two main components: first, a structure recognition algorithm which visually tracks pixel regions captured by its camera; second, an action recognition algorithm which classifies the physical actions of intelligent entities (such as humans or robots). Objects in the environment are recognized by associating an observed action with a particular environmental feature. As an example, by knowing where chairs are, a robot can know best where to expect humans to spend most of their time while working in an office setting. By observing a human sitting down, the FOCUS algorithm can classify the object on which the human sat as functionally a chair. This would include any other object where a human would sit, including small tables, couches, heat registers, or even boxes. Thus, the problem space is reduced from needing to reason about multiple object types to the problem of motion recognition and classification which can be robust across different environments.

By finding one object in the image, we can then generalize and find multiple similar objects. We aim at giving the found examples to a learning algorithm, which will be hopefully capable of finding increasingly better general descriptions of objects. FOCUS uniquely contributes to the field of functional object recognition and learning through observation in the following ways. First, FOCUS does not require specific visual models of the environment or the objects presented within it to be known ahead of time. This is in contrast to other object recognition algorithms which require some prior or exemplar knowledge. A robot's sensory model of an object is blank until an observation is made of a human that interacts with the object in a known fashion. Secondly, the object descriptor that FOCUS employs abstracts away from any specific low-level feature detection modalities. Whatever feature detection is employed by FOCUS merely requires a method for applying the algorithm onto an image and a method for comparing the parameters of different features of the same type to determine a measure of similarity.

The paper is organized as follows. Section 2 discussed some related work. Section 5 describes the complete object recognition algorithm while sections 3 and 4 discuss the lower-level visual feature extractors as well as the human activity recognition algorithms, respectively. Section 6 describes empirical validation of the method and section 7 concludes the paper.

2. RELATED WORK

Representing a concept as an aggregate of multiple different modes of thinking has been proposed by [23]. In the representation of FOCUS, we consider both the affordance property of the object as well as the sensory definition and merge these two representational modes to create a much richer description of the object.

Our approach relies on visual structural information to extract candidate features from the scene. Some examples of this range from active contour models [18, 4], to autonomous vehicle guidance on a highway [9], and tracking a leaf in a cluttered background [17]. Other work has been devoted to the problem of learning specific features through methods such as PCA [24]. In [25], context is used to help disambiguate objects observed in scenes. These methods worked well because of their use of appropriate prior knowledge. However, they required much more complete visual object models to be known *a priori* and our work assumes that the robot must learn such knowledge exclusively by observing humans.

Our approach also relies on human activity recognition which is a topic that has been studied fairly extensively for a number of years. In [13], hand gestures such as writing in different languages on a whiteboard, as well as activities such as using a Glucose monitor are recognized. In [16], finite state machine models of gestures are constructed that by learning the spatial and temporal information of the gestures separately from each other. An extension to the Hidden Markov Model [28] formalism called the Abstract Hidden Markov mEmory Model (AHMEM) [26] is used to represent both state-dependent and context-free behaviors. A system for torso and arm position tracking is described in [8] which fits *a priori* geometric models to data collected from a stereo camera. Classifying the behaviors of individual (non-human) agents using hidden Markov models has been studied in [14] where an overhead camera is used to track soccer playing robots. However, much of the research has relied on very rigid assumptions about the placement of the sensor with respect to the object being tracked. Our work does not rely on such assumptions and can be considered to be more general.

Functional object recognition has also been examined from the standpoint of using visual recognition of the parts of an object to determine how that object can be used [32]. As an example, a tool such as a hammer can be recognized by first identifying some sort of striking head as well as a handle [29]. Other approaches include reasoning about the causal structure of physics in a scene [6]. A similar approach is called Form and Function [36]. Gibson's theory of affordances [12] can be applied to this sort of definition in that certain objects have forms that are derived from their functions. Thus, the specific shape of an object can be used to identify how the object can be used and thus ultimately identify the object itself. One example is the ratio of height to width of individual stairs in a staircase [22]. Another example can be found in the relative spatial locations of an object's components [5]. Thus can be derived one of the key ideas that inspires our research: the affordance of an object is directly related to its function. Our approach makes no assumption about the structure of an object, nor of the physical models that may govern it but rather uses an observation of a human activity that interacts with that object to identify it as belonging to a specific class. Only then are the structural features of that object learned and generalized. However, those features are only useful from the standpoint of being able to recognize similar features in other objects.

Additionally, activity tracking was used in concert with known object models and context in order recognize a variety of different objects [7]. The difference between this approach and our approach is that image-based object models are known ahead of time. No concept of how an object is used is incorporated into the disambiguation process. Our approach does not make use of known object models but rather segments the image based on the detected activity and updates the object model accordingly.

In work very closely related to our own, [15] combines a structural hierarchy of objects and empty spaces (voids) with tracked activities of people interacting with them to identify objects by their use. In this work, objects are classified as either vertical or horizontal supportable objects on which someone can place or hang portable objects. Additionally, voids are recognized as doorways when someone moves through them. A combination of a passive video camera and depth maps from stereo are used to pre-segment the image in order to identify these physical characteristics. Our approach only requires a single video camera to run and also only focuses primarily on identifying the objects based on the specific set of activities that the human performs on them rather than requiring an ontology to describe the space. Our approach also includes a generalization phase which takes the learned visual characteristics of a particular object and finds other objects like it in the environment without having to observe a person interacting with them first.

Training a robot about its environment and the objects within it has been examined from the standpoint of direct human intervention, cooperative learning, and robotic self-discovery. For instance, in [35] a system of gesture-based programming was developed based on a multi-agent model of "encapsulated expertise." Robot to robot action recognition and cooperation has been successfully accomplished using a stereo vision system[20] and extended by [2] to "learning by imitation" in which a robot learns its behaviors by observing the behaviors of another robot. A system that learned the parameters for teleoperated manipulations was developed by [27]. The paradigm of learning from observation has also been used very successfully to train humanoid robots on specific tasks [10] as well as action generation [3]. Object recognition and learning object models from multiple sources and modalities of information has been explored by [1]. Another method, described in [11] and again in [30], discusses how a robot can use active perception (vision and tactile information) to explore its environment to learn structural properties of objects through manipulation. Our algorithm draws from these two paradigms by making use of multiple sources of information gleaned by observing human activities to learn the visual object definitions.

3. CANDIDATE FEATURE SEGMENTATION

An object to be identified has a structural and functional definition as well as spatial and temporal connections which relate the activity recognizer with the structural recognizer. Table 1 lists the object features and gives a concrete example of a chair, as we have implemented.

Object Features	Example: Chair
Structural	Colored region
Functional	Sit still(face) returns the face position as Frame:F_t,pixel:P_s
Spatial Connection	PixelDist(P_s) returns Δp, s.t. chair is at $P_{s-\Delta p}$
Temporal Connection	PastFrameDist(F_t) returns Δf, s.t. chair is at $F_{t-\Delta f}$

Table 1: Chair as structure and function

The structure feature is defined in terms of color, meaning that the model says that a chair has a color but does not specify which color. Color as an object feature, as opposed to a CAD model, is computationally inexpensive and more environmentally invariant. Clearly, color by itself does not identify a chair, but will do so when combined with the activity recognizer. The activity "Sit still" is recognized by a face at some pixel location and at a frame at some time. Table 1 shows the spatial and temporal connections between an activity of "Sit still" and a chair. The spatial connection returns a pixel displacement to be applied to the position of the face in order to search for the chair. Because the body should occlude the chair, there is a temporal connection between the body and the chair. This temporal connection shows which frame the chair should be visible in the past. Spatial and temporal connections can either be given explicitly in the model, or could be associated with procedures to determine them, i.e. the temporal connection procedure could return the first frame where the person is not occluding an object.

Because each low-level feature abstraction must define its own specific candidate feature and similarity measure components, we present example algorithms for the contiguous region feature. This feature detector uses color and shape homogeneity constraints to segments each image into contiguous regions of similar color. Currently, two different types of low-level features are implemented in FOCUS. The first is a contiguous region tracker, described in [34], and the second is based on the PCA-SIFT [19] algorithm. The former algorithm, illustrated in Figure 2, tracks regions from frame to frame and thus does not have to re-segment each new image. Each of these tracked regions becomes potential candidate features for objects.

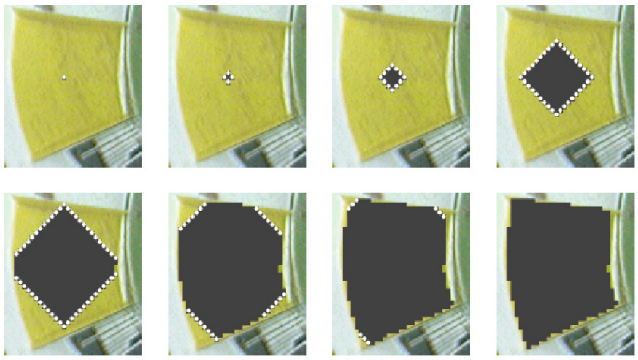

Figure 2: Example of the region growing algorithm starting in the upper left and going to the lower right. In this figure, the boundaries for light-colored region are found by an expanding boundary illustrated by white dots. The final region is shown in black.

4. ACTIVITY RECOGNITION

We assume that the objects we want to recognize are used by humans and therefore, we want to detect those human's activities. We detect people by searching a video stream for faces [31]. Because natural human environments contain objects that are typically unknown and unmodeled, detecting people's bodies is very challenging and difficult to do properly. As a result, the most robust feature for detecting a person becomes their face. Faces are very unique and distinguishing features which greatly simplify the task of determining whether an object is a person or not. The face detection algorithms that we use are a compromise between the need for accurate and robust person detection and the needs for CAMEO to operate in real-time. The decision to process facial data only directly affects the selection of specific features for activity recognition.

The relative displacements of the face positions are calculated (difference from one frame to the next) and given as input to the activity recognition algorithm. We have defined a finite state dia-

gram that encapsulates a coarse level of possible activities that can be detected in the data stream of visually tracked faces. Figure 3 illustrates the state diagram of the activities FOCUS can now recognize.

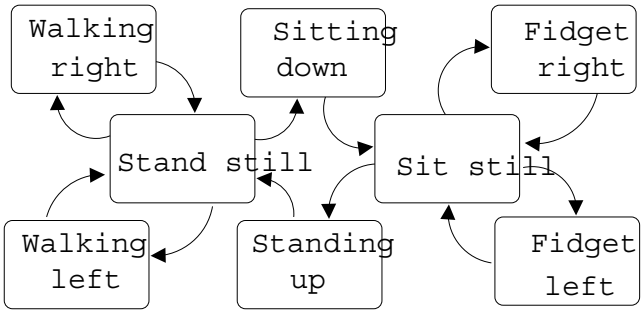

Figure 3: Example finite state diagram for a single person in a meeting.

The "Sit still" state represents instances when a person is sitting in their chair and is stationary. The "Fidget left" and "Fidget right" states represent motion of those people in their chair, such as if they look around or lean to one side to talk to someone. The "Standing up" and "Sitting down" states represent the transitions from a "Sit still" and "Stand still" state. These are the actual activities involved with getting up from one's chair and taking one's seat, respectively. Finally, once a person is in the "Stand still" state, they can "Walk left" and "Walk right".

We use the Viterbi algorithm [28] to infer the person's state given the sequence of data from the detected face positions. Viterbi is a dynamic programming algorithm which takes into account the state transition probabilities to generate the most likely state sequence that could have generated the observation sequence. This is more formally defined as:

$$q_{1:t} = arg\max_{q_1:t} P(q_{1:x}|y_{1:t}) \quad (1)$$

where q_t is the most likely state at time t. Thus, for each timestep, Viterbi computes a term $\delta_t(j)$, defined as:

$$\delta_t(j) = \max_i \left[\delta_{t-1}(i)P(X_t = j|X_{t-1} = i)\right] P(y_t|X_t = j) \quad (2)$$

which is initialized as:

$$\delta_0(i) = P(X_0 = i)P(y_0|X_0 = i) \quad (3)$$

Additionally, the index of the most likely state at time $t-1$ to transition into each state at time t is computed and stored in a table $\psi_t(j)$ as follows:

$$\psi_t(j) = arg\max_i \left[\delta_{t-1}(i)P(X_t = j|X_{t-1} = i)\right] \quad (4)$$

Finally, after the entire observation sequence has been analyzed in this fashion, the state sequence is obtained by backtracking over the table of ψ values ranging from $t = 0$ to $t = T$:

$$q_T = arg\max_i \left[\delta_T(i)\right]$$
$$q_t = \psi_{t+1}(q_{t+1}) \quad (5)$$

After the most likely activity state sequence has been computed for each tracked face, the location of the face is returned by the action recognition algorithm along with the specific activity name. This is used by FOCUS to identify when a human has potentially used an object of interest.

5. OBJECT CLASSIFICATION BY USE AND STRUCTURE

FOCUS combines activity recognition with visual feature extraction to learn to associate what features belong to a specific functional object type. Figure 4 illustrates the overall FOCUS approach. In addition to the face tracking and activity recognition modules for identifying people and their activities, FOCUS uses a feature extraction and segmentation module to identify potentially significant structures in the environment. While in this example, the sensory data is provided by a stationary camera, the algorithm can easily be generalized by mounting the camera on a moving robotic platform and tracking the objects from frame to frame as the robot moves in its environment. The structural feature detector used by FOCUS (described in section 3) was originally implemented for tracking contiguous regions in the environment by a fast-moving robot.

5.1 Object Class Definitions

FOCUS connects the output from the activity recognizer with the list of potential object candidates returned by the visual feature segmentation algorithm. This allows the algorithm to infer the functional relationship between objects and activities and associate a specific set of visual features with the object in question. FOCUS object class definitions consist of several components, some of which must be defined *a priori*. The pre-defined components represent the prior knowledge about humans and activities that must be known before the visual features of any object can be learned. The unknown component of an object is the visual feature definition which, unlike the other components, will be learned dynamically.

Affordance property Each object class to be detected must have a specific activity that can be recognized when the human uses it. For FOCUS, a chair is any object that a human can sit down upon, and a door is any region that a human can walk through. By identifying the associated actions when they occur, the location of the person in the image is used to specify low-level visual features that are part of the object.

Spatial association Just as with the activity recognition module, several *a priori* assumptions are made about the relationships of people and the environment. Portals will be larger than the person walking through them, "chairs" will be below head-height, "tables" will not be taller than a person, etc... Heuristics can be employed to identify typical regions that are the walls and ceiling (if visible). These regions can then also be excluded from consideration. When an activity is detected, the low-level visual feature that best matches the spatial association criteria is selected as the best candidate.

Temporal association When an action is detected, the person interacting with the object will typically occlude the object from the robot's point of view. Therefore, FOCUS cannot assume that it can find the object in the current image using the spatial information inferred in the previous section. Instead, a previous image from a stored history must be used. In this model, the index of the previous frame can be given or can be inferred by assuming a maximal size of the object. By searching backwards over the face tracking results, an image can be found when the person's location was not close to their current position. If the robot has never viewed the area behind the person, the algorithm will fail and the robot must either wait for the person to get up and move, or it should move to a new view point and wait for a similar action to take place later.

Figure 4: Flow diagram of FOCUS. In this example, an omni-directional camera captures raw video from the environment. This video is searched for human faces. When a face is found, the person is tracked over time. The motions of the individual faces are analyzed for specific activities. Finally, the "functional" aspects of the activity are used to classify "structural" features of objects used in that activity (i.e., chairs are identified where people sit).

Visual descriptor Initially, this component of an object class is empty as the robot has not observed any data that can be stored within it. When a specific visual feature is associated with a detected human activity, the details of the visual feature descriptor along with its parameters are stored in this field of the object class. If distinctly different feature types are detected later but which still correspond to the same object type (such as the different chairs in Figure 1), they are appended to the list of possible feature descriptors. Thus, for this example, FOCUS can greatly enrich its definition of objects that can function as chairs.

Object generalization Having associated a candidate feature for the chair object, FOCUS searches the image for an instance of similar features. A match would be an indication of potential other chairs. In order to accomplish this, a similarity measure is defined which allows FOCUS to compare feature sets found in the regions. In our example, region features include mean color, size, and boundary shape. For the case of color, we define two regions to be similar if their color distance is below a given threshold t_c based on the Euclidean color distance in RGB color space. Similarly, their boundary shapes must match based on comparison of their curvature parameters.

5.2 The Algorithm

The FOCUS algorithm tracks humans as they move around their environment. The structural feature processor segments the image into chunks based on homogeneity constraints defined by the visual feature extraction algorithm. When an activity is recognized, that activity is used to find the object class whose specific affordance property matches. The spatial correspondence and temporal correspondence features of that object class are then used to identify the specific visual structural component in the image that will be correlated with the activity. This object's specific visual features are stored in the visual descriptor and the parameters are used to perform a search for similar objects in the image that match those parameters based on the object generalization parameters.

6. EMPIRICAL VALIDATION

Figure 5 shows an illustrative example of the FOCUS algorithm operating on a video stream. A chair is defined as an object with some color (not specified ahead of which color.) Two different types of chairs, red and blue, are recognized through two persons sitting on one chair of each type. The region segmentation algorithm is also running over this image and is identifying contiguous regions of similar color (boundaries between regions are shown as black lines). In Figure 5(a), a person is tracked and is identified as walking while in Figure 5(b), the person is identified as standing. In Figure 5(c) the person is identified as sitting down. Using the knowledge about the spatial relationship between the position of the face and the expected location of the chair, the segmented region which is closest to the expected position of the chair is considered to be a structural feature of that chair. However, FOCUS needed to return to an image before Figure 5(b) to find that candidate region as the person moved and occluded the chair from that point on. To find hypotheses for other chairs, the system searches the image for regions with similar color (see Figure 5(c)). Nearly all of the red chairs were found in this way. However, since there were two types of chairs with two different colors (red and blue) present in the room, not all chairs are found. In Figure 5(d), a second person walks through the room and sits down now in a blue chair. In this instance, the same algorithm is performed but now, a blue blob is identified as a chair. All regions that are similar to this one are identified as being chairs of type 2, as shown in Figure 5(e). From this example, we can see that FOCUS has identified an association between a particular set of low-level vision features and the activity of sitting down. The results of this particular experiment are shown in the table below:

Chair type	Total	Detected	False positives
Red	15	9	3
Blue	5	3	0

The table shows the total number of chairs of each color type that could have been detected is compared to the actual number that were detected. Additionally, any incorrectly objects incorrectly labeled during the detection phase are also shown. Note that not all

of the chairs in the room are identified properly. Partially, this is because not all of the chairs are necessarily the same shape as the exemplar chair from the robot's point of view. The shape differences are too great with respect to the similarity metric that is used for the color/shape detection algorithm. These chairs would be detected once another person sat down in them and the other objects of similar shape/color would then be labeled appropriately. This is also an example of how one structural feature detector (shape and color) can be used. FOCUS is agnostic to the specifics of the feature detector and can make use of any other sort of structural feature detector algorithm such as PCA-SIFT [19] or KLT [21] [33] tracked features. The benefit for a robot that uses FOCUS to understand its environment is the association of environmental features with particular human affordance properties. Even if the robot is incapable of directly manipulating the object (for example, how can a wheeled robot sit down?), it is sill capable of reasoning about that environmental feature as an object that must be taken into account when reasoning about human interactions.

7. CONCLUSIONS

Object recognition is a well-known and difficult problem, particularly from images where different views of objects and taken in different environments. In this paper, we have contributed an approach that aims at overcoming two main issues: (i) the brittleness of object recognition with respect to signals in specific environments; (ii) the difficulty of accurately defining universal structural properties of objects. FOCUS combines a simple structural definition of objects with their use. Object usage is tracked in terms of motion detection which is quite robust with respect to the environment. FOCUS tracks the use of an object through recognizing activities of people in interacting with the environment. A simple structural definition, in terms of general features is associated with the activity when the latter is detected. FOCUS then generalizes to finding the candidate similar objects in the image. For robots that must interact with humans in natural environments, the space of possible object definitions can be very large. However, by defining objects rather by a smaller set of possible human uses, this space can be reduced to a tractable size. Additionally, by reasoning about how the objects are used by humans, a robot that makes use of the FOCUS algorithm take on a much more human-centric perspective which in turn can lead to more natural human/robot interactions. Our method shows that a poorly-defined structural model of an object combined with its use can be powerful at classifying and objects in a scene. The paper presents the specific fully-implemented example of identifying a chair. The method extends to any objects by adding the corresponding structural and activity models.

FOCUS continually learns by observing how humans directly interact with objects. Right now, there exists no notion of identifying objects *indirectly* within FOCUS, but this is an area of active future research. An important distinction between FOCUS and other functional object recognition approaches is that FOCUS makes no restrictions on the kinds of objects that can belong to the different classes. That is, classes do not need to be disjoint sets. For example, a table could belong to a class of "chairs" (things to sit on) as well as a class of "ladders" (things to climb on.) This acknowledges the simple fact that a single object can be multi-function. Because the learned entries for these classes are data-driven, higher-level descriptions (such as formal dictionary definitions of objects) do not bias the learning algorithm. As a result, FOCUS is capable of learning how objects can be used in novel fashions that fit outside their "traditional" uses and definitions.

8. ACKNOWLEDGMENTS

This research was supported by the National Business Center (NBC) of the Department of the Interior (DOI) under a subcontract from SRI International. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, by the NBC, DOI, SRI or the US Government.

9. REFERENCES

[1] Artur M. Arsenio. Object recognition from multiple percepts. In *IEEE-RAS/RSJ International Conference on Humanoid Robots*, Los Angeles, USA, November 2004.

[2] P. Bakker and Y. Kuniyoshi. Robot see, robot do: An overview of robot imitation. In *AISB Workshop on Learning in Robots and Animals*, Brighton, UK, April 1996.

[3] Darrin Bentivegna, CHristopher Atkeson, and Gordon Cheng. Learning from observation and practice at the action generation level. In *IEEE International Conference on Humanoid Robots*, Karlsruhe and Munich, Germany, September/October 2003.

[4] A. Blake and M. Isard. *Active Contours*. Springer-Verlag, 1998.

[5] A. M. Borghi. Object concepts and action: Extracting affordances from object parts. *Acta Psychologica*, 115:69–96, 2004.

[6] Matthew Brand. Physics-based visual understanding. *Computer Vision and Image Understanding: CVIU*, 65(2):192–205, 1997.

[7] Darnell J. Moore and Irfan A. Essa and Monson H. Hayes III. Exploiting human actions and object context for recognition tasks. In *Proceedings of the International Conference on Computer Vision, V1*, pages 80–86, 1999.

[8] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. *International Journal of Computer Vision*, 37(2):175–185, June 2000.

[9] E. D. Dickmanns and B. D. Mysliwetz. Recursive 3-d road and relative ego-state recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 14:199–213, February 1992.

[10] M. Ehrenmass, R. Zöllner, O. Rogalla, S. Vacek, and R. Dillmann. Observation in programming by demonstration: Training and execution environment. In *IEEE International Conference on Humanoid Robots*, Karlsruhe and Munich, Germany, September/October 2003.

[11] Paul Fitzpatrick. Object lesson: discovering and learning to recognize objects. In *IEEE International Conference on Humanoid Robots*, Karlsruhe and Munich, Germany, September/October 2003.

[12] James Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, New Jersey, USA, 1979.

[13] Raffay Hamid, Yan Huang, and Irfan Essa. ARGMode – activity recognition using graphical models. In *Conference on Computer Vision and Pattern Recognition Workshop*, volume 4, pages 38–44, Madison, WI, June 2003.

[14] Kwun Han and Manuela Veloso. Automated robot behavior recognition applied to robotic soccer. In John Hollerbach and Dan Koditschek, editors, *Robotics Research: the Ninth International Symposium*, pages 199–204. Springer-Verlag,

London, 2000. Also in the Proceedings of IJCAI-99 Workshop on Team Behaviors and Plan Recognition.

[15] Mirai Higuchi, Shigeki Aoki, Atsuhiro Kojima, and Kunio Kukunaga. Scene recognition based on relationship between human actions and objects. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 73–78, 2004.

[16] Pengyu Hong, Matthew Turk, and Thomas S. Huang. Gesture modeling and recognition using finite state machines. In *Proceedings of the Fourth IEEE International Conference and Gesture Recognition*, Grenoble, France, 2000.

[17] M. Isard and A. Blake. Condensation: conditional density propagation for visual tracking. *International journal of Computer Vision*, 1998.

[18] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Proc. of IEEE Conference on Computer Vision*, pages 259–268, 8-11 1987.

[19] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, volume 2, pages 506–513, 2004.

[20] Y. Kuniyoshi. Vision-based behaviors for multi-robot cooperation. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, volume 2, pages 925–932, Munich, 1994.

[21] Bruce D. Lucas and Takeo. Kanade. An iterative image registration technique with an application to stereo vision. *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[22] L. S. Mark. Eye height-scaled information about affordances: A study of sitting and stair climbing. *Journal of Experimental Psychology: Human Perception and Performance*, 13:361–370, 1987.

[23] Marvin Minsky. *Society of Mind*. Simon & Schuster, March 1988.

[24] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *International Conference on Computer Vision (ICCV'95)*, pages 786–793, Cambridge, USA, June 1995.

[25] Kevin Murphy, Antonio Torralba, and William Freeman. Using the forest to see the trees: A graphical model relating features, objects and scenes. In *NIPS'03, Neural Information Processing Systems*, 2003.

[26] N. Nguyen, H. Bui, S. Venkatesh, and G. West. Recognizing and monitoring high level behaviours in complex spatial environments. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2003.

[27] P.K. Pook and D.H. Ballard. Recognizing teleoperated manipulations. In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 2, pages 578–585, 1993.

[28] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[29] Ehud Rivlin, Sven J. Dickinson, and Azriel Rosenfeld. Recognition by functional parts. *Computer Vision and Image Understanding: CVIU*, 62(2):164–176, 1995.

[30] Ryo Rukano, Yasuo Kuniyoshi, Takumi Kobayashi, and Takuya Otani. Statistical manipulation learning of unknown objects by a multi-fingered robot hand. In *Proceedings of the International Conference on Humanoid Robotics*, 2004.

[31] Paul E. Rybski, Fernando de la Torre, Raju Patil, Carlos Vallespi, Manuela M. Veloso, and Brett Browning. Cameo: The camera assisted meeting event observer. In *International Conference on Robotics and Automation*, New Orleans, April 2004.

[32] L. Stark and K. Bowyer. Generic recognition through qualitative reasoning about 3-d shape and object function. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–256, Maui, HI, 1991.

[33] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. Technical report, School of Computer Science, Carnegie Mellon University, April 1991.

[34] Felix von Hundelshausen and Raul Rojas. Tracking regions and edges by shrinking and growing. In *Proceedings of the RoboCup 2003 International Symposium, Padova, Italy*, 2003.

[35] R. M. Voyles and P.K. Khosla. Gesture-based programming: A preliminary demo. In *Proceedings of the IEEE International Conference on Robotics and Automation*, Detroit, 1999.

[36] Kevin Woods, Diane Cook, Lawrence Hall, Kevin W. Bowyer, and Louise Stark. Learning membership functions in a function-based object recognition system. *Journal of Artificial Intelligence Research*, 3:187–222, 1995.

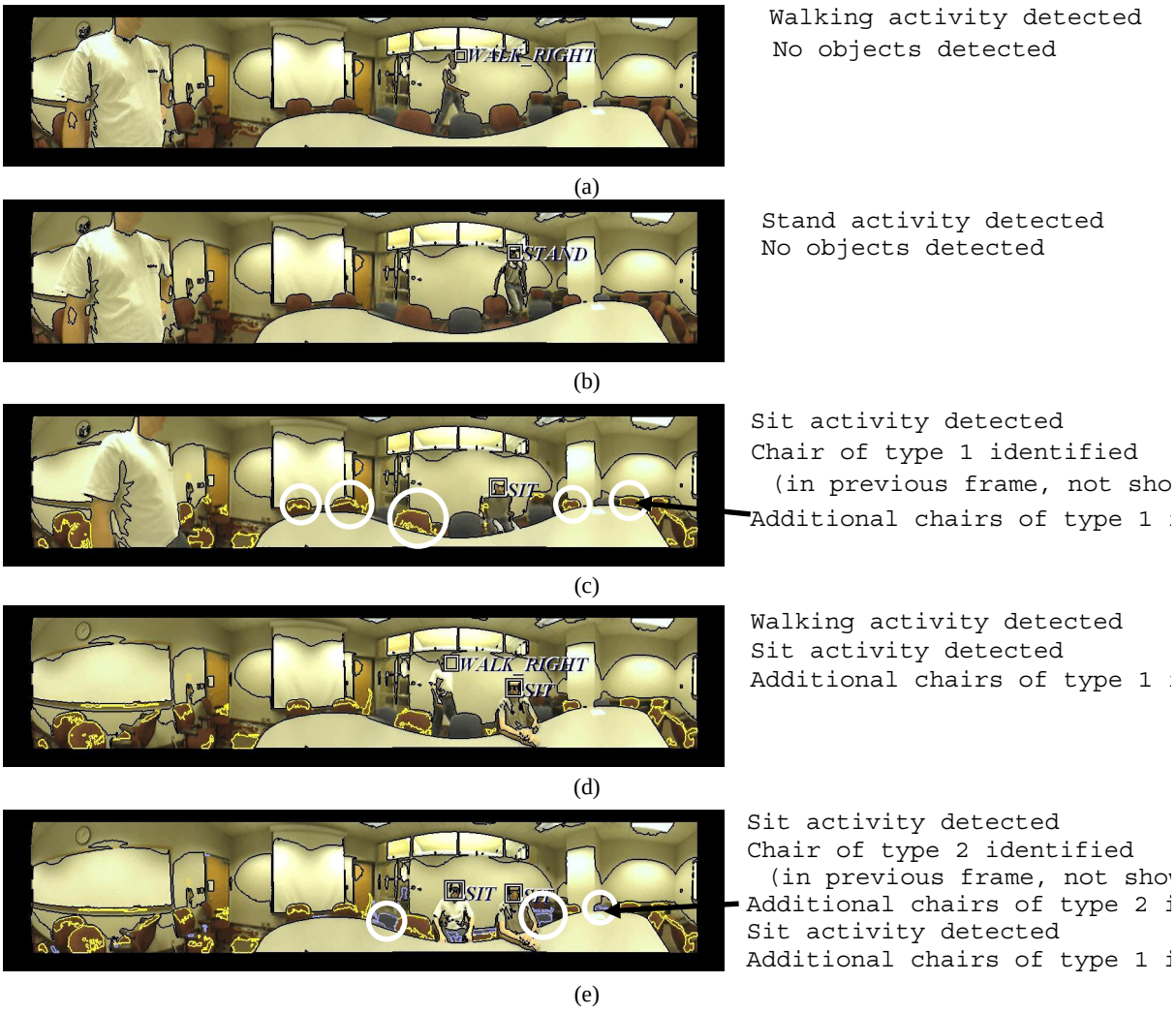

Figure 5: Snapshots from a video illustrating FOCUS running. In (a)-(b) a human's face is tracked while walking to a chair. At the same time the image is segmented into homogeneous regions. In (c) the action recognition algorithm determines that the person is now sitting in a chair, and the segmented region, associated in space and in a previous frame, instantiates the visual characteristics for the chair, in this case colored red. All regions with a similar color are detected, in order to find chairs of the same type. In (d)-(e) the process is repeated with a second person, sitting on a second type of chair.