Context-Dependent Multi-Cue Object Recognition

Sarah Aboutalib
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
saboutal@cs.cmu.edu

December 16, 2008

Proposal Addendum

Committee:
Manuela Veloso (Chair)
Martial Hebert
Alexei Efros
Irfan Essa (Georgia Institute of Technology)

1 Thesis Approach: A Clarification

The ultimate goal of my thesis work is to further the path towards achieving the fundamental definition of object recognition for computer vision, that is, having a computer determine the identity of a set of objects and their location in an image or a set of images. More specifically, an object is determined to be recognized if the correct label is attached to a point that lies somewhere within the true segmentation of the object in the image. This point represents the localization of the object. It is not necessary to retrieve the true segmentation as defined in our problem, but it is a bonus if possible. Thus, with this definition of *object recognition*, we move on to a discussion of possible approaches to achieve this goal.

The problem of general object recognition has come to be associated with a particular technique of using mathematical models and signal processing to determine complex representations and exact comparisons of models and extracted features from images (such as the use of CAD models that try to capture the exact 3D geometry of each object). These methods view the object recognition task as a direct mapping of an object model to objects in an image using mathematical transforms to account for alterations in orientation, scale, rotations and other adjustments. While these methods may be good for finding objects in constrained environments, where very specific objects in very specific backgrounds need to be exactly localized for manipulation or inspection, these techniques do not fair well once taken out of those environments and can provide poor recognition results under generic real world conditions that are very much unconstrained. In order to deal with the vast variation in real world data, it would be necessary to create one of these intricate models for every possible appearance and form of each object class, a very daunting task.

However, there is another approach to the core object recognition challenge, the approach I would like to explore with my thesis and that is utilizing numerous simple fast feature extraction methods and cues and the *combination* of all this data to distinguish one object from another. This approach opens up new applications of these algorithms since we are focused on fast and easily accessible cues, we can apply object recognition in a real way to real moving data, with robust solutions to real problems. Although our model could in fact use the more complicated model representations, it is not necessary especially if it gets in the way of recognizing in real-time real video data. Because our algorithm prefers fast simple information, it is then necessary to broaden the scope of the type of information it can use, i.e., we want to expand past just the visual feature of the object to include easily retrieved activity, speech, sound, and any other type of information. Thus, we have provided an algorithm and representation that can indeed combine multiple types of information for the recognition of these objects.

Part of the type of information we would like to utilize for this approach in object recognition falls under functional recognition, that is, the recognition of objects based on their functional properties and affordance cues. For instance, recognizing a chair as something someone can sit upon, or looking for handle-shape parts of an object (see Functional Recognition Section for additional descriptions) to determine if it can be picked-up (i.e., a pickable object). This, however, is not the only type of information we would like to utilize in recognizing an object. In conjunction with function, we would also like to use other types of cue information such as the mentioning of the object during a conversation, (i.e., speech information), which may have nothing to do with its function (For instance, 'that chair looks very comfortable'). In addition, we are not interested in simply representing objects in terms of their functions, placing objects of equivalent function under the same object category. We want to be able to recognize chalkboards separate from paper, even though both are objects for writing. We want to label objects according to the classic categorization including its physical structure and visual features to distinguish them as well as information outside of these two realms. Thus, you can view multiple-cue object recognition as a broader category of object recognition than either functional or visual-based.

In order to allow for the inclusion of all these various types of information, providing an evidence gathering technique is an important part of my thesis work. It is not however limited to this topic of evidence gathering, as there are already systems and approaches out there that provide flexible information gathering techniques for any type of application. My thesis includes taking these approaches such as Probabilistic Relational Models and applying to images and the problem of object recognition. Thus, we don't simply want to determine the presence of an object given evidence, we actually *want to find it in an image.* This requires very specific image related information and manipulation. For instance, the importance of the cue representation to include temporal and spatial information so that the correct region is segmented in an image, and although a simple color segmentation approach may not lead to the true segmentation of an object, it is enough to locate a point on the object and to place the correct object label in the image, which is what we defined as object recognition in the opening paragraph. Furthermore, our algorithm provides the unique ability to generalize new learned cues to find additional objects within the image.

In summary, we want to use image-based visual descriptions, in conjunction with functional information, in conjunction with other types of cue information to solve the ultimate goal of object recognition. We place our thesis work under the category of object recognition albeit with an approach that expands the set of tools beyond those typically used for this problem.

2 Functional Recognition

Now because an important part of the information utilized by my approach does rely heavily on functional information and the use of the objects. It is important to have an understanding of functional recognition and approaches in this area.

Figure 1: Classic object categories (left) vs. functional object categories (right) [7]

The first step is to give a definition of functional recognition: Functional recognition is the recognition of objects based on their function, i.e. a functional definition categorizes and defines objects according to their potential uses and their utility in performing a task [7]. In figure 1, the differences between a functional and a classic categorization of objects is shown. Note that a functional category may put together two objects that would normally be categorized separately as in the mug and the water-can and classic object categories may include several function categories. Thus, if one wanted to give the objects classic object recognition labels, one could not do so on the basis of a functional definition alone.

The functional definition of object categories depends on the notion of affordances introduced by [4, 3]. In Gibsons theory of affordances, certain aspects of an objects form can be used to determine their function. Affordances were 'action possibilities' defined in relation to the actor and thus, the actors capabilities. So, if we are talking about a human, a door knob affords the ability to be turned because of its form, i.e. its a turnable object for that human. (For a bird on the other hand, the door knob would not provide the same affordance since its form does not lend to the birds ability to turn it. However, we will stick with the perspective of an adult human as that makes most sense in terms of most object recognition problems). Thus, affordance properties are how an object can be interfaced with, defining the objects utility. For object recognition, this is a useful concept as it allows the approach to use affordance properties or cues to determine the function of an object. For example, in some works affordance properties are defined as the relative spatial locations of an object's components [2].

Functional recognition then uses these affordance properties to determine the function and finally the identity of the object. In most functional recognition approaches, the interaction of a human with an object is used in order to identify in the object either in place of or in conjunction with its visual attributes. In many cases, visual recognition techniques are used first to recognize parts of an object in order to determine its use (and thus functional object category) [7, 5, 12], with one of the first being [11], who used visual properties in order to determine geometric properties to inform the function of the objects, such as the bent (arch) nature of a handle on a cup or watering-can, which allows for grasping. In some cases, the affordance properties are used first in order to identify objects from which visual properties are taken to find additional objects [9]. But under almost all cases, an object is labeled according to its functional definition, not the classical.

3 Evaluation

In this section, I will be describing the methods of evaluation I will be using in my thesis work. I will be going over three main components: (1) the sources I will be using to retrieve cue information from the datasets, (2) a description of proposed datasets I will be using to evaluate my thesis contributions including which cue sources I will be using, and (3) the concepts I wish to evaluate and the measures I will be using to determine success as well which datasets will be used in order to so.

3.1 Sources of Cue Information

We have numbered here all the possible different sources of cue information I will be using in my thesis work. These numbers will be used in the Dataset section to indicated which cue sources are used by each dataset. The sources here are divided into three sections: Automatic, By Hand, and In-Between.

3.1.1 Automatic

This section provides a description of the automated recognition systems that I will be using to generate cues for my thesis work. By automated, we simply have to give the recognition system the video from which we would like to run our algorithm and it will generate the cues without any information being fed by hand. For each cue source, we describe the exact output generated including what cue information is given, how it's given, when it's given and what information is used by the the MCOR algorithm.

(1) Rybski et al Activity Recognition This is an activity recognition system developed by [6], which can label the activities of multiple people in a scene, as long as their faces are pointed towards the camera. At each frame, the system outputs one of six activity labels (Sit, Stand, Fidget_Left, Fidget_Right, Walk_Left, Walk_Right) for each person at a given pixel value (determined by the location of the persons face)[6].

Cue Information Used by MCOR:

From this system, the MCOR algorithm can extract Activity Cues, i.e., for every frame, we get an activity label associated with each person in the scene and a pixel value representing the location of the label.

(2) Contiguous Region Tracker This code was taken from Felix von Hundelshausen and segments a region based on a region growing algorithm [10] that finds contiguous regions of similar color. The segments are then tracked from image to image so a re-segmentation is not needed. In addition, we will only be looking at segments with other cue information attached so the entire scene.

Cue Information Used by MCOR:

At every frame for each segment, the color and aspect ratio (height over width) of the bounding box is given. This gives color and shape information. Often times, the color and shape information are tied together when looking for objects of the same visual features, such that only segments of a particular color and within a small difference of the aspect ratio will be considered.

It is also possible to learn PCA-SIFT features from these regions as an additional visual feature as mentioned in [9].

(3) Qrio Activity Recognition This is an activity recognition system I developed for the Qrio video data (It can possibly be adapted to other datasets in the future, although currently it is just for the Qrio data). It recognizes small (Pick_Up, Put_Down, Stay, Move_Right, Move_Left) and large movements (Writing, Eating, and Talking)[1].

Cue Information Used by MCOR:

This recognizer generates an activity cue. For each frame, it returns an activity label of either small or large movements as specified and a pixel location attached to the center of a colored wristband around one hand of the person that is doing the activity.

3.1.2 By Hand

This is cue information which I will have to input myself by hand. But, please note, at any point in time, it is possible to replace any of the hand information by an automated system.

(4) Speech Information This gives cue information in imitation of speech recognition systems. At base levels, the set of words that can be returned will correspond to the object labels of the objects to be recognized. It is possible for particular datasets to include such phrases as 'look that up' in order to identify a laptop. These speech cues will be clearly indicated when results are given.

Cue Information Used by MCOR: For each frame, a phrase or word is returned along with a pixel location representing the source of the speech (such as the mouth of a person as located by hand).

(5) Sound Information This gives cue information in imitation of sound recognition systems. More specifically, we will primarily be using cues that could be recognized by the UPC AED/C (Acoustic Event Detection and Classification) System [8]. Thus, labels representing 12 possible classes of sounds including Knock, Door Slam, Steps, Keyboard Typing, and Phone Ringing will be returned.

Cue Information Used by MCOR: For each frame, a label indicating one of the 12 sound classes will be returned along with a pixel location representing the source of the sound (such as the center point of the phone that is ringing, which will be located by hand).

(6) More activity information This adds any additional activity information that falls out of the scope of the two activity recognition systems described in cue sources (1,3). This includes activities difficult for current fast activity recognition algorithms, such as 'putting_objects_down'. This is also useful for simulated data in which the typical video data would not be available for the activity recognition systems. Currently, the set of possible activities will be: Sit, Stand, Erasing, Pointing, Put_Objs_down, Talking, Writing, Pick_Up, Drinking, and Eating.

Cue Information Used by MCOR: For each frame, one of the above activity labels is returned along with a pixel location representing the location of the activity.

3.1.3 In-Between

This is cue information that is retrieved from already annotated video with written cue and label information, that I do not need to input by hand myself, but which was not generated by an automatic recognition system either. This information is usually specific to the dataset.

(7) Instructions This is text containing step-by-step instructions describing what is happening in each scene. For example, if cooking data was being used, this cue source would come in the form of a recipe, with each step of the recipe identified with a particular segment of the video.

Cue Information Used by MCOR: For all frames under a particular segment step in the instruction, a number of cue information can be provided such as activity, speech, and other objects (For instance, if the instruction says 'stir the eggs in the bowl', you then can extract the activity 'stir', the speech cue 'bowl' (assuming the instructions are being said out loud), and the object 'eggs'. The location the cues will have to be identified by hand.

3.2 Datasets

The datasets can be split along two dimensions: (1) my data vs datasets from outside, and (2) simple vs complex. In the sections below, I will first describe my datasets for use in evaluation. In the second section, I will describe the other datasets. Within each section, I will move from more simple datasets to complex. For each dataset, I will list the cue sources that will be used in each dataset as defined in the previous cue source section:

3.2.1 My Data

***1* Simulation** This is a simulator described in detail in the proposal.

INPUT: The simulator is given a text file containing the set of objects in the scene and their location at each time frame. Locations remain the same unless an activity that causes the object to move is applied. This is specified by another file describing the behavior of the activity. In previous work, the simulator did not cause any objects to be moved, but this aspect can be taken advantage of in the future. In addition, a file containing all the probabilities of a each cue being produced given the presence of an object, and a file representing the object dictionary and all the cue properties (including spatial and temporal association) as described in the proposal.

OUTPUT: It generates Speech, Activity, Color, Shape, and Sound cues based on the given set of objects and probabilities. The output is represented by a text file that identifies the frame number and the list of cues generated for that frame. For each of these cues, the cue type, the cue value, and its location is given. The location is determined by the location of the object it was generated from plus the spatial association found in the object dictionary. In order to mimic the imperfection of the real world, noise was added to the data generated by the simulator according to the error rate of the recognition systems that the cues were based on, whenever applicable. In other words, if it is known that a system has .9 accuracy, 10 percent of the time the simulator will produce an incorrect cue value.

Because this simulator generates its own cues, it does not use any of the cue sources described above, rather it bases its cues according to those found in the object dictionary given at the outset.

Objects: Whiteboard, Projector Screen, Table, Laptop, Chair, Pen, Paper, Bottle, Plate, Fork, and any number of objects can be easily added to this list.

***2* Same Room, One Person, Scripted** This dataset includes all video in which a single person is recorded interacting with the environment according to a given script with all the video taken from a single context or room.

Some of this data has already been collected and can be summarized as thus:

Figure 2: Sample images from dataset *2*. The image on the left is from the CAMEO device. The image on the right is from the Qrio data.

First, we have data from a CAMEO device, that is, a set of four or five firewire cameras arranged in such a way to capture a full panoramic view of the room. The images from these cameras are merged together into a consistent mosaic producing video with a frame size of 1085x260 pixels). Three videos were taken which were about half a minute to a minute long. Each were recorded in a Newell-Simon Conference room.

The following sources of cue information were used: (1,2,4,6)
Objects: Whiteboard, Projector Screen, Table, Laptop, Chair

In addition, there is data from the Qrio camera, which consisted of two videos about a minute long each with frame sizes of about 176x144 pixels. They were filmed in the CORAL Lab on the first floor of wean. The robot remained static during the recording.

Cue Sources: (2,3,4)
Objects: Pen, Fork, Paper, Cellphone, Plate

***3* Different Rooms, One Person, Scripted** This dataset includes all video in which a single person is recorded interacting with the environment according to a given script with video taken from a number of contexts, i.e. different rooms.

Some of this data has already been collected and can be summarized as thus:

I collected data using a Panasonic Digital Camera mounted on a tripod with frame sizes of 704x480 pixels. Video was taken from 5 contexts: NSH conference room, NSH classroom, NSH atrium, Wean classroom, Wean Conference Room. For each context about 2 to 3 videos were taken of about 2 minutes each. (so at least 20 [2x2x5] minutes of video).

Cue Sources: (1,2,4,5,6)
Objects: Whiteboard, Projector Screen, Table, Laptop, Chair, Chalkboard, Projector, Bottle, Pen, Paper, Phone, Eraser, Clock

***4* Different Rooms, Multiple People, Scripted** This dataset includes all video in which a multiple people are recorded interacting with the environment according to a given script with video taken from a number of contexts, i.e. different rooms.

This data will be collected using a Panasonic Digital Camera mounted on a tripod with frame sizes of 704x480 pixels. Video will be taken in at least 3 contexts (such as the NSH Conference room, Wean Classroom, and NSH Atrium). There will be about 2-3 people in each video following the script.

Cue Sources: (1,2,4,5,6)
Objects: Whiteboard, Projector Screen, Table, Laptop, Chair, Chalkboard, Projector, Bottle, Pen, Paper, Phone, Eraser, Clock

***5* Different Rooms, Multiple People, Unscripted** This dataset includes all video in which a multiple people are recorded interacting with the environment in a real scenario (no script) with video taken from a number of contexts, i.e. different rooms.

I will ask people at CMU in real meetings and possibly lectures to record with a Panasonic digital video camera. The locations will depend on approval, but at least two different contexts will be given.

Cue Sources: (1,2,4,5,6)
Objects: Whiteboard, Projector Screen, Table, Laptop, Chair, Chalkboard, Projector, Bottle, Pen, Paper, Phone, Eraser, Clock

3.2.2 Other Data

The rest of the evaluation will be done on data from sources outside.

***6* Simulated Video Data** We will use the simulated video data software provided by ObjectVideo Virtual Video (OVVV).

The OVVV Tool generates realistic video from simulated cameras in an interactive virtual world. This tool is free and is based on a modification (aka 'mod') of Half-Life 2, a commercially available game from Valve Software.

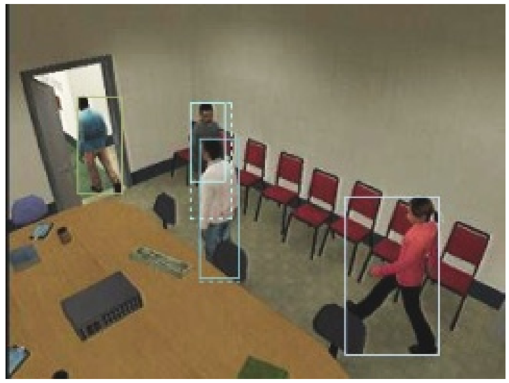

Figure 3: Sample image from dataset *6*

The simulator allows you to create maps that include furniture and objects that the simulated people interact with using the videogame AI (as described in the manual). In this way, an exact segmentation of each object would be known, and the environment can be set up.

Realistic virtual video requires realistic content in the form of maps, models, animations and script-based scenarios. There is sample content in the Source SDK which includes a lot of the objects used in my own data (chairs, tables, etc) as well as many others for very versatile training (see http://developer.valvesoftware.com/wiki/List_of_entities for a full list of the possible entities and objects). In addition, I can also create custom maps and models to capture specific scenarios of interest. Also, camera noise can be added to the data in order to simulate more real world circumstances.

http://www.objectvideo.net/index.html

Cue Sources: (2,4,5,6, possibly 1, in addition, to the color, shape, and segmentation provided by the virtual nature of the world)

***7* Larry Davis Data** Larry Davis and his group have a nice dataset of subjects performing 6 interactions with 4 objects. The objects include cup, spray bottle, phone and flashlight. The interactions with these objects were: drinking from a cup, spraying from a spray bottle, answering a phone call, making a phone call, pouring from a cup and lighting the flashlight. There are also other objects, such as a stapler, which the Davis group did not train for, but which we can use for our own thesis work.

The advantage of the Larry Davis set is that it gives us object recognition video which we can test against. The disadvantage however, is that the data is primarily taken on an almost solid background behind the people and objects, with not much occlusion or real world background. Thus, further tests in less constrained real world environments is necessary (see next dataset).

http://www.umiacs.umd.edu/l̃sd/

Cue Sources: (1,2,4,5,6)

***8* eHow.com** With all the previous datasets within this category had very controlled environments with fairly solid and plain backgrounds for the objects and people. We would then like to test our algorithm on

Figure 4: Sample image from dataset *7*

real world data without regard for such control. Our plan is to use the eHow.com video set.

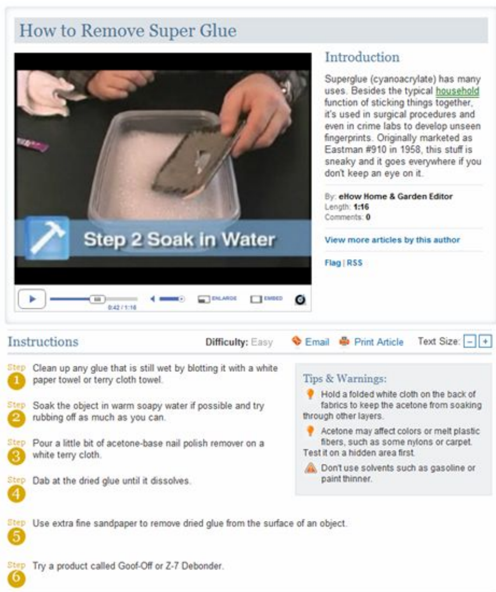

Figure 5: Sample image from dataset *8*

These are videos with attached instructions for each step. We plan to use the eHow.com website that provides step-by-step instructions on how to do a variety of activities. This is useful because it provides a large range of potential objects and settings, along with already labeled activity, speech, other objects, and other possible cues in the corresponding instruction data.

Cue Sources: (1,2,4,5,7)

3.3 Concepts For Evaluation

There are four major contributions that my thesis work would like to provide (see figure 6 for summary):

Generalization This is the ability of my algorithm to learn new object traits (i.e., add additional cues to

		Datasets	Generalize	Weights	Context	Object Rec.
Simple	Mine	*1* Simulation		X		X
		2 Same, Single,Script	X		X	X
		3 Diff, Single, Script	X		X	X
		4 Diff, Multiple, Script	X		X	X
Complex		*5* Diff, Multiple,Unscrip	X		X	X
Simple	Other	*6* Virtual Simulation		X		X
		7 Larry Davis				X
Complex		*8* eHow.com			X	X

Figure 6: Summary of the concepts I wish to evaluate in my thesis (top row) and the datasets I wish to use for this evaluation (3rd column from the left). An 'X' is placed for each dataset that will be used to evaluate that concept. Datasets are separated into two categories: Mine and other, in reference to the datasets that will be created by myself, and those that will be taken from outside sources. In addition, under each category, the data sets are ordered from simple to complex.

that object's definition) after it has been recognized. In other words, to generalize the traits of the recognized object to all objects in that category, so that additional objects can be found based on this new information.

This concept will be evaluated by determining whether any objects were recognized based on information that was not found in the initial definition. This concept will be determined successful if even one additional object was found on this base. Of course, the more objects found, the better the technique.

For evaluation, we will use datasets *2,3,4,5*, i.e. all datasets except the simulation data (since the definitions for these objects are usually completely defined at the outset) and Larry Davis Data set since there are usually not a large number of the same type of object.

Weight Learning This is the ability of my algorithm to learn the strength of the association between a particular object and a particular cue (as described in detail in the proposal).

For this, we will compare the learned weights with the original weight model in order to validate the PRM learning mechanism. In addition, we will test the recognition using these weights on a test set.

Since learning requires a large amount of data, we will use the simulation datasets *1,6*.

Context Specific vs General This is the ability of the algorithm to learn object definitions specific to a particular context as well as learn what properties make the general definition of the object regardless of context.

In order to test, the algorithms ability to adapt to different contexts, I will need to use datasets with a variety of environments. Thus, I will datasets *2,3,4,5* and possibly *8*.

Object Recognition This is the ability of the algorithm to identify and locate objects in a scene.

This is the ultimate test of my thesis approach as it ties in all the above areas together and address directly the ultimate goal of my thesis work as described in the beginning of this document.

In conjunction with our definition of object recognition described in the Thesis Approach section, we will evaluate the success of our algorithm by determining the number of objects correctly recognized within a particular scene. This means by the end of the video, we will count the number of objects in which a point of identification was placed by the algorithm with the correct object label and within the true segmentation of the object (as determined by hand or simulation; this is ground truth). All objects, which do not contain such a point or contain the wrong label, will be counted as incorrect. Thus, a comparison of the number of objects correctly identified in the scene can be determined.

Note: Only objects found in the object dictionary, i.e. specified by the user for recognition will be counted.

All the datasets can be used for evaluation using this method, i.e. datasets: *1,2,3,4,5,6,7,8*.

References

[1] Sarah Aboutalib and Manuela Veloso. Cue-based equivalence classes and incremental discrimination for object recognition. *International Conference on Robotics and Automation (ICRA)*, 2009. under submission.

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

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

[4] J.J. Gibson. *The theory of affordance. In: Percieving, Acting, and Knowing.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.

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

[6] Paul E. Rybski and Manuela M. Veloso. Robust real-time human activity recognition from tracked face displacements. *In Proceedings of EPIA'05, the 12th Portuguese Conference on Artificial Intelligence*, 2005.

[7] M. Stark, P. Lies, M. Zillich, J. Wyatt, and B. Schiele. Functional object class detection based on learned affordance cues. *Computer Vision Systems: 6th International Conference (ICVS)*, pages 435–444, 2008.

[8] A. Temko and C. Nadeu. Classification of meeting-room acoustic events with support vector machines and variable-feature-set clustering. *In IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, 5:505–508, 2005.

[9] Manuela M. Veloso, Paul E. Rybski, and Felix von Hundelshausen. Focus: a generalized method for object discovery for robots that observe and interact with humans. In *Proceedings of the 2006 Conference on Human-Robot Interaction*, March 2006.

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

[11] P.H. Winston, B. Katz, T.O. Binford, and M.R. Lowry. Learning physical descriptions from functional definitions, examples, and precedents. *AAAI*, 1983.

[12] 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.