Towards Using Multiple Cues for Robust Object Recognition

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

Manuela Veloso
Carnegie Mellon University
Computer Science Department
Pittsburgh, Pennsylvania
veloso@cmu.edu

ABSTRACT

A robot's ability to assist humans in a variety of tasks, e.g. in search and rescue or in a household, heavily depends on the robot's reliable recognition of the objects in the environment. Numerous approaches attempt to recognize objects based only on the robot's vision. However, the same type of object can have very different visual appearances, such as shape, size, pose, and color. Although such approaches are widely studied with relative success, the general object recognition task still remains very challenging. We build our work upon the fact that robots can observe humans interacting with the objects in their environment, and thus providing numerous non-visual cues to those objects' identities. We research on a flexible object recognition approach which can use *any multiple cues*, whether they are visual cues intrinsic to the object or provided by observation of a human. We realize the challenging issue that multiple cues can have different weight in their association with an object definition and need to be taken into account during recognition. In this paper, we contribute a probabilistic relational representation of the cue weights and an object recognition algorithm that can flexibly combine multiple cues of any type to robustly recognize objects. We show illustrative results of our implemented approach using visual, activity, gesture, and speech cues, provided by machine or human, to recognize objects more robustly than when using only a single cue.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Perceptual reasoning, Video analysis

General Terms

Algorithms

Keywords

Computer Vision, Object Recognition, Multiple Cues

AAMAS'07 May 14–18 2007, Honolulu, Hawai'i, USA.

1. INTRODUCTION

There are a variety of tasks where the robot agent's ability to assist humans depends heavily on the reliable recognition of the objects in the environment. Object recognition however has proven to be a significantly difficult challenge especially with the complexity of real world data, where there is great variation in both the appearance of objects within a single class e.g. chairs come in many shapes and colors, and in the appearance of the same object under various circumstances, e.g. the same chair can appear different with changes in lighting, view, orientation, etc.

A number of approaches have attempted to focus on learning the visual features of an object (see 'related work') in order to recognize it. Although great progress has been made along these lines, there is still much to be done in order to build an object recognition system that can be run under any of the various situations that must be dealt with with real data.

In dealing with this complexity, an important observation is to note that as humans interact with their environment, they are providing numerous non-visual cues to the identity of objects within it which can be utilized by a robot observing the interaction. The benefit of including non-visual information is supported both by the success made by numerous approaches which have integrated non-visual cues such as activities in their system and by biological studies of the human visual system described further in the 'related works' section.

Our approach is then to provide a flexible framework for the inclusion of *any multiple cues*, whether they are visual cues intrinsic to the object or provided by observations of a human for more robust object recognition. This framework includes three main components:

First is a generic representation of the cue information. The standardization of the information provided by the various cues allows for the evidence of cues of any type–activity, speech, vision, gesture, etc.– to be taken into account without special modification to the object recognition algorithm. This allows the algorithm to deal with real world scenarios where the type of cues that are available will vary: sensors may be broken, outside circumstances may prevent correct readings, different agents may have different capabilities. The inclusion of different cue information is important since it is not just visual cues in particular that may have weaknesses when it comes to object recognition, but any cue in isolation. For instance, with speech cues, although there are a number of instances where its linguistic content can be used to disambiguate objects in a picture [9], it is far from a perfected science and chances of false positives or failures in a real-world situation is not unexpected. If one were to use activity cues alone, there would also be numerous possibilities of error such as how to deal with an activity that could describe more than one object or false positives caused by an activity that misused an object. Other types of cues have similar weaknesses.

Second is the inclusion of weights for the cue evidence. We recognize the interesting and challenging fact that some cues may be more indicative of an object than others and thus the evidence given by that cue should have greater influence. In order to reflect this fact, weights are added whose value represent the strength of the association between a particular cue and object for each possible cue and object. The value of the weights are determined using probabilistic techniques that learn the strength of the association through training data.

Third is an algorithm, MCOR (Multiple Cue Object Recognition), which utilizes the representation of cues and weights to determine the evidence for the presence of an object and generalizes new cues found in the recently recognized object to all objects of the same class i.e. the object definition is changed to include the new information.

Empirical validation of the multiple cue framework is given through illustrative results of the implemented approach using visual, activity, gesture, and speech cues.

2. RELATED WORK

As mentioned above, numerous approaches to object recognition such as [4, 1, 11, 7] depend on visual cues alone. Although great strides have been made using this technique, demonstrating fast and high accuracy results, the performance of these approaches and most visual-based object recognition systems to some extent, have often been subject to the effect of variations in such aspects as size, lighting, rotation, or pose prevalent in most real world images. Some approaches such as [3] attempt to counter some of these concerns by finding features of an object invariant to those changes. Although this does show robust performance to variations in size and rotation, it only partially handles lighting and pose. Current research is still looking into further refinement of recognition by visual cues in order to improve on these weaknesses.

Other techniques however attempt another approach of including cues that do not depend on the visual properties of the object itself to compensate for the weaknesses of the visual cues. For instance, in a paper by Murphy et al.[6], scene context is used to provide additional evidence for the presence of an object. So, for example, if one were in an office, one could expect to see a computer screen or keyboard and so the scene context, the office, can provide evidence for those two objects in addition to their particular visual features. This, however, lends a way for incorrect classification or a missed recognition caused by an object being placed in a scene it normally is not in. Others such as [5] attempt to add a completely non-visual cue, i.e. activities, in addition to the visual properties. Similar to [5], the FOCUS algorithm [10] is based on visual features and activity cues, but unlike it, it does not need any training examples for the visual description ([12] also illustrates a method of recognition without human labeled training examples, although in terms of activity recognition not object).

In addition to the relative success of these approaches, the inclusion of non-visual information for more robust object recognition is further supported by biological studies of the human visual system where, with little effort, most humans have the ability to recognize the same object even when there are great variations. Laboratory studies have shown however that this robust ability to recognize objects becomes limited when a person is shown only an image of the object where much of the context is removed and the object is in a non-canonical position (so that previous experience with the object can not be utilized in the recognition), as shown by an increase in response time [8] when the person attempts to name it. Context, i.e. other non-visual cues associated with the object, allows the person to be confident that the object he/she is holding is a teapot, even though the visual cues of the object itself may be lacking when looked at in isolation.

Supported by the overall success of these approaches in integrating a non-visual cue for more robust object recognition and by biological findings, our approach provides a general framework for flexibly including multiple cues of any number and any type, so that all the cues mentioned above such as activities, visual features, and context, in addition to any other possible cues available now or in the future, can be used to provide evidence for the presences of the object.

3. REPRESENTATION

3.1 Object Dictionary

In previous object recognition methods, an object is usually described using particular types of information defined at the outset, e.g. a graph of visual features [7] or, as in the FOCUS algorithm [10], a functional and visual description, and, although the content of the information is allowed to change, the actual structure of the definition is not. In this paper, we break the necessity of having to define and therefore limit the type of information that can belong to an object. We do so by defining an object as a set of cues, C_{o_i} (o_i represents the i^{th} object to be recognized), where the cues can be of any type and the set can be of any size, as long as each cue follows a standard format described below.

Using this definition, we can then describe an algorithm (see 'Multiple Cue Object Recognition' section for details) that is independent of the type and number of cues available and thus can utilize the wide variety and varying cues humans provide in their interaction with the object. Segmented regions in the image can then be recognized as objects by comparing the cues extracted from the scene with cues in the object definition using the properties defined below.

3.2 Cues

In order to integrate the evidence from multiple sources i.e. from multiple cues, it is necessary to have a standard representation of the information. Thus, all cues must define a set of properties:

cue type, *cue_type***,** the kind of information provided by the cue, e.g. activity information, speech information, visual, etc.

cue value, *cue_value***,** the output of the extraction method of that particular cue type, e.g. an activity value, such as SITTING, extracted by an activity recognizer for

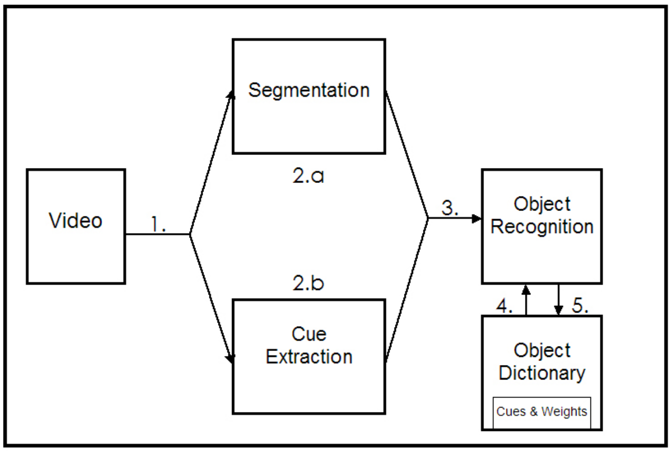


Figure 1: Flow diagram of MCOR framework: (1.) Get image from video, (2.a) segment the image, while at the same time (2.b) extract cues from the image, then (3.) associate extracted cues with a segment, (4.) recognize objects based on dictionary, and (5.) update object dictionary based on the recognized objects.

activity cues, a word or phrase extracted by a speech recognizer.

spatial association, Δp**,** the difference between the current cue's position (a cue-specific calculation)and the expected location of the object to which the cue is associated, e.g. if the location of the activity, SITTING, was defined as the center of the face of the person sitting, then the spatial association would be the distance between the center point of the face and the expected position of the chair. This property will be used to determine which segmented region in the scene the cue belongs to.

temporal association, Δf**,** the difference between the current frame and the frame where the object is clearly visible. This is primarily important for cues that in the process of being produced might obscure the object they indicate, e.g. the activity of SITTING may obscure the chair that it is providing evidence for. A clear view of the object is necessary in order for it to be segmented and recognized.

weight, w_{o_i}**,** the strength of the association between an object, o_i, and the cue.

similarity measure, the method for calculating the similarity, s_{c_j}, between this cue and another c_j, e.g. the euclidean distance in RGB color space for visual cues based on color. If the cue, c_j is of a different type then $s_{c_j} = 0$;

It is possible for there to be multiple cues of the same type in the same definition. It is not allowed, however, to have cues with the same *cue_value*, since duplicate cues do not provide additional information to the description of an object.

4. WEIGHTS

As mentioned above, there are weights ascribed to each cue in each object definition. These weights represent the strength of the association between that cue and the object. This allows for the fact that different cues may be more indicative of an object than others.

The weight is then determined by the probability of the object o_i being present given the cue value, *cue_value*, i.e. $P(o_i|cue\_value)$.

By increasing or decreasing the value, a greater or lesser dependence on the cue in the calculation of evidence for a particular object label can be enacted.

4.1 Probabilistic Relational Models

Although a number of probabilistic techniques could have been used to calculate the values of the weights, Probabilistic Relational Models (PRM) [2] was chosen since it is an extension of Bayesian networks that can include the relational information of data. Where Bayes net apply only to flat, i.e. attribute-value [2], representations of the data, PRM can learn associations between classes, attributes within a classes, and attributes related to another class. Thus, PRM allows for future growth in learning weights that reflect the relationship of the various properties of the cues and objects in determining the weights, although at this particular time only a simple weight representation is utilized.

The relational model of a PRM is defined by a schema which consists of several components: First is the set of classes, $\mathbf{X} = X_1, ..., X_n$ where each has a set of attributes, $A(X_i) = X_i.a_1, ..., X_i.a_2$. The attributes can be either 'fixed' or 'probabilistic'. 'Fixed' attributes are there to identify instances of the class (referred to as *entities*) and thus their value does not change. The value of 'probabilistic' attributes however can vary based on the other attributes of the entity or of related entities. It is this affect that we attempt to model and learn the parameters of.

In our case, there are two classes: cues and objects, i.e. $\mathbf{X} = CUE, OBJECT$. The attributes of the CUE class consists of the cue identity and the cue value, i.e. $A(CUE) = cue\_id, cue\_value$ and the object class consists of the object identity and the object label, i.e. $A(OBJECT) = obj\_id, obj\_label$, where the identity in each case is a fixed attribute and the *cue_value* and *object_label* are probabilistic.

The second component is the set of relations, $\mathbf{R} = R_1, ..., R_m$ which defines the relationship between two classes. Relationships are significant in that the value of attributes in one class can depend not only on the other attributes of that class, but on the attributes of any related class. In our case, we define a single relationship, $\mathbf{R} = ASSOCIATED\_WITH$, which links an object and cue class whenever the cue is associated with that object.

It is possible with PRMs to learn the dependency structure, S, between the attributes of the classes, but since in our case the schema is simple, we can assume there is a dependency by the label of an object on the cue value, i.e. $P(OBJECT.obj\_label|CUE.cue\_value)$, which would make up the parameters of the dependency structure, δ_S

PRMs then describe a probability model over instances of a relational schema. An instance, I, of the relational schema consists of the set of entities of each class,$O^\sigma(X_i) = e_1, ..., e_p$, where the attributes are defined and which relationships exist between them. A skeleton, σ, is when only the fixed attributes of the entities are defined. In our case, an instance would consist of all the objects in a scene, all the cues in the scene, and the association between any of the cues or objects.

Some of the attributes however are not easily defined such as the object label and thus it is necessary to determine the probability distribution of its values. A relational skeleton σ is then a partial instance where the probabilistic attributes are undefined.

PRM then defines the distribution of instantiations of attributes as:

$$P(I|\sigma, S, \delta_S) = \prod_{X_i \in \mathbf{X}} \prod_{a_j \in A(X_i)} \prod_{e_k \in O^{\sigma}(X_i)} P(I_{e_k.a_j}|I_{pa(e_k.a_j)}) \tag{1}$$

where in our particular case, $X_i \in CUE, OBJECT$, $a_j \in cue - value$ or $obj - label$ (the fixed attributes are ignored, since it would not make sense to learn the probability of their value), and $e_k \in O^{\sigma}(X_i)$ is the partial instantiation of each entity, i.e. the objects with unspecified object labels. $pa(e_k.a_j)$ are the parents of the j^{th} attribute in the k^{th} entity.

Given a training set, the parameters δ_S can be learned according to the following equation:

$$\begin{aligned} l(\delta_S|I,\sigma,S) &= logP(I|\sigma,S,\delta_S) \\ &= \textstyle\sum_{X_i} \sum_{A \in A(X_i)} \left[\sum_{x \in O^{\sigma}(X_i)} logP(I_{x.a}|I_{pa(x.a)})\right] \end{aligned} \tag{2}$$

Standard maximum likelihood estimation can then be applied where δ is chosen in order to maximize l.

5. MULTIPLE-CUE OBJECT RECOGNITION

The algorithm for object recognition with multiple cues, i.e. Multiple Cue Object Recognition (MCOR), then proceeds as follows (see figure 2 for pseudocode of the algorithm.):

Given a set of cues (also referred to as the object definition), C_{o_i}, associated with an object, o_i, for each object in the set of objects to be recognized, O, and a video, we can begin object recognition:

For each frame of the video, F_t. The first step is to extract all cues which belong to the union of the set of all cues in all of the object definitions, i.e. $\bigcup_i C_{o_i}$, in addition, which type of cues are extracted is dependent upon which tools are available. For instance, if an activity recognizer is implemented, it could be used to extract activity cues. If a speech recognizer is implemented, speech cues can be extracted, and so on. That is the major benefit of this algorithm: it does not require any specific type of extraction method, but rather utilizes whatever is available. It is this characteristic which allows the algorithm to be run by any robot or computer under any circumstance. In addition, it is the various recognizers that will track and extract cues from the interaction of the human with the object.

5.1 Region Extraction

Processing then continues with each new cue extracted, where a new cue is defined as a cue that was not present in the previous frame, F_{t-1}, with the same *cue_value* and at a position P_j less than a cue-specific distance away. Then, for each new cue, c_j, the current position, P_j will be retrieved. This is also a cue-specific calculation.

If the predicted location of the object belonging to that cue, i.e. $P_j - \Delta p_j$ is within a region, r_k, from the set of segmented regions to be recognized as objects, R, that cue, c_j, is stored in a set of cues that belong to that region, r_k, i.e. C_k. If, however, $P_j - \Delta p_j$ does not fall within an already segmented region, a new region will be segmented at that location, but from frame $F_{t-\Delta f_j}$, where the object is clearly visible. This new region is then stored in R.

```
Given a set of cues for each object:

• For each object, $o_i$, in the set of possible objects to be recognized, $O$:
    – There should be a set of cues, $C_{o_i}$.
    – Each cue, $c_l$, in $C_{o_i}$ represents a cue that is associated (i.e. indicates) object $o_i$ and which has:
        ∗ a cue value, $cue\_value_l$
        ∗ a temporal association, $\Delta f_l$
        ∗ a spatial association, $\Delta p_l$
        ∗ a weight, $w_{o_i,c_l}$
        ∗ a similarity measure to calculate the similarity, $s_{c_j,c_l}$ between cue $c_l$ and another cue $c_j$

Analyze the video:

• For each frame of the video, $F_t$:
    – Extract all cues that belong to $\bigcup_i C_{o_i}$
    – For each new cue extracted, $c_j$, with $cue\_value_j$:
        ∗ Get current position, $P_j$.
        ∗ If $P_j - \Delta p_j$ at $F_{t-\Delta f_j}$ is within any region $r_k \in R$, where $R$ is the set of segmented regions to be recognized as objects:
            · Store $c_j$ in $C_k$, the set of cues attached to that region.
        ∗ Else:
            · Extract a new region, $r_k$, at position $P_j - \Delta p_j$ and frame $F_{t-\Delta f_j}$ and store it in $R$.
            · Store $c_j$ in the currently empty $C_k$
    – For each region, $r_k \in R$:
        ∗ For each object, $o_i \in O$:
            · Calculate the evidence, $e_{k,o_i}$, that region $r_k$ is object $o_i$ as follows:
                  $e_{k,o_i} = \sum_{c_l \in C_{o_i}} \sum_{c_j \in C_k} w_{o_i,c_l} s_{c_j,c_l}$
              if the cue type of $c_l$ is not the same as $c_j$, then $s_{c_j,c_l} = 0$
        ∗ Region $r_k$ is then recognized as the object with the greatest evidence, if it is above a threshold, $\theta$, i.e.
                  $label_k \leftarrow \text{argmax}_{o_i} e_{k,o_i}$, if $\max e_{k,o_i} > \theta$
        ∗ Add all cues, $c_j \in C_k$, to the set of cues in the object definition, $C_{label_k}$, if $\forall c_l \in C_{label_k}, s_{c_j,c_l} \neq 1$
        ∗ If the current label, i.e. $label_k$ at $F_t$ is different from $label_k$ at $F_{t-1}$ and $label_k$ at $F_{t-1}$ exists:
            · Remove all $c_j \in C_k$ added before $F_t$ from $C_{o_{old}}$, where $o_{old} = label_k$ at $F_{t-1}$.
```

Figure 2: Algorithm for Multiple-Cue Object Recognition

5.2 Calculation of Evidence

For each region, r_k, in R, and for each object, o_i, in O, the algorithm calculates the evidence, e_{k,o_i}, that r_k should be labeled o_i according to the following equation:

$$e_{k,o_i} = \sum_{c_l \in C_{o_i}} \sum_{c_j \in C_k} w_{o_i,c_l} s_{c_j,c_l}$$

where w_{o_i,c_l}, as described earlier in the 'Weights' section, is the predefined weight representing the strength of the association between the cue, c_l, in the object definition, C_{o_i}, and the object, o_i. s_{c_j,c_l} is the similarity between the cue c_l and the cue, c_j, belonging to C_k. If the cues are not of the same type, then $s_{c_j,cue_l} = 0$.

5.3 Object Recognition

Region, r_k, is then recognized as the object with the greatest evidence, if it is above a given threshold, θ, i.e.,

$$label_k \leftarrow \text{argmax}_{o_i}\, e_{k,o_i}, \text{ if } \max e_{k,o_i} > \theta$$

5.4 Generalization

All the cues, c_j, in the set, C_k, of cues belonging to the region, r_k, will now be added to the object definition of o_{label_k}, i.e. C_{label_k}, if there is no cue in the definition with the same *cue_value*, since duplicate cues are not allowed, i.e. $\forall c_l \in C_{label_k}, s_{c_j,c_l} \neq 1$. This generalizes the cues learned from this particular object to all objects of the same class, so now new objects based on the newly added attributes can be found.

If the algorithm were to stop here, however, there would be a problem if the region, r_k, had been previously labeled as another object, i.e. $label_k$ at $F_{t-1} \neq label_k$ at F_t. (For simplicity, we will call the old object label, $label_k$ at F_{t-1}, o_{old}.) This is because all the cues belonging to that region, except those that were newly extracted at the current frame, would have been added to o_{old}'s definition with the previous iteration, when we now can assume, since with the addition of more evidence the object label changed, that was the wrong object label, and so the cues of that region should not belong to o_{old}'s definition. Thus, it is necessary to remove any cues that were added to r_k before the current time step from the definition of o_{old}, i.e. remove all cues, $c_j \in C_k$ added before F_t from $C_{o_{old}}$.

6. EXPERIMENTS AND RESULTS

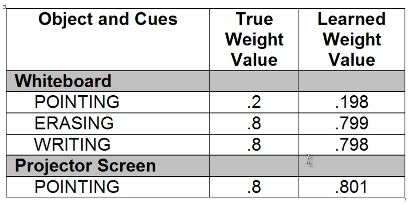
6.1 Weight Learning Using Simulation Data

In order to illustrate how the weight values for each cue and object could be learned, synthetic data was generated by a simulator. Given a set of objects, the simulator generated cues based on predetermined model which represents the probabilities of a cue being produced given the presence of an object. It is this model which the weights attempt to learn using the PRM technique outlined above.

In order to demonstrate how the weights used in the real data scenarios described below could have been learned, the simulator was set up using those predefined weights (see 'Empirical Validation using Real Data') as the model. Thus, three scenarios were produced by the simulator matching the three scenarios of the real data. The weights generated by the PRM learning are then compared with the true model values as shown in figures 3, 4, 5. The simulation generated 100 runs for each scenario in order to learn the weights.

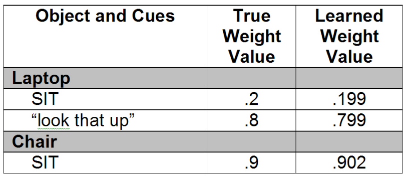
In the first scenario, there were two objects, a whiteboard and a projector screen, where the related cues, i.e. POINTING, ERASING, WRITING, and their probabilities are represented in 5 in addition to the learned weights. In the second scenario, there were two objects, laptop and chair. The cues consisted of a speech and an activity cue, i.e. 'look that up' and SIT. The third scenario consisted of a table and chair with the cues being the actions PUT_DOWN and SIT.

With an average error of .003 between the true and the learned weights, one can see that the PRM learning technique was able to successfully learn the true model used by the simulator.



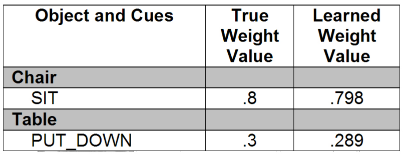
Object and Cues	True Weight Value	Learned Weight Value
Whiteboard		
POINTING	.2	.198
ERASING	.8	.799
WRITING	.8	.798
Projector Screen		
POINTING	.8	.801

Figure 3: Comparison of learned weights to true weights in first scenario.



Object and Cues	True Weight Value	Learned Weight Value
Laptop		
SIT	.2	.199
"look that up"	.8	.799
Chair		
SIT	.9	.902

Figure 4: Comparison of learned weights to true weights in second scenario.



Object and Cues	True Weight Value	Learned Weight Value
Chair		
SIT	.8	.798
Table		
PUT_DOWN	.3	.289

Figure 5: Comparison of learned weights to true weights in third scenario.

6.2 Empirical Validation using Real Data

These results illustrate the feasibility of the MCOR framework. Three object recognition tasks were given to test various capabilities of the algorithm: the first demonstrates how the algorithm can deal with an unreliable cue type, the second demonstrates how the algorithm resolves cases where the same cue indicates the presence of different objects, and the third demonstrates how the algorithm deals with a misleading cue.

All the tasks started off with object definitions which contained cues that could be observed from the interaction of a human with the object and a weight value corresponding to the strength of the association between the cue and the object (currently human determined). The definitions are begun with cues that involve human interaction since those cues tend to be more obviously associated with the object since humans usually interact with the object in a manner implicit to its definition. For instance, most interactions tend to portray the function of the object which is usually an important part of its definition. Additional cues are learned by the algorithm later (see figure 6(b,e) and figure 7(b,d)) according to the generalization method described earlier.

In addition, for all recognition tasks, segmentation is based on a color-based region growing algorithm, also described in detail in [10]. The similarity measure for the activity and visual cues will be binary, where $s_{c_j,c_l} = 1$, if the cues are the same, 0 otherwise, where 'same' for the activity cue means that the activity label is the same for each cue and 'same' for the visual cue means the color distance between the cues and the shape ratio as defined in [10] is within a threshold, t_c. This means that the value of the evidence will simply depend on the sum of the weights of the cues found in the definition, thus for simplicity, we will not refer to the value or multiplication of s_{c_j,c_l} in the sum, although it is implicitly there. Cue values were extracted either by external recognition and/or vision systems such as the color and shape extraction defined in [10] or by hand-encoded values such as the activity and speech cues, since the purpose of this paper is to illustrate how various cues can be combined the actual source of the cue is not of very much importance.

6.3 Unreliable Cue Type

For the first object recognition task, the goal was to recognize two objects: a whiteboard and a projector screen, which tend to have very similar visual features, in this case, similar color and shape–a task which most visual based object recognition systems would have extreme difficulty with. In both cases, however, MCOR was able to successfully label each object (see figure 6(e)). Figure 6(a) illustrates how a new cue, a visual *color&shape* cue, is added to the definition of the projector screen. The figure shows each key time step as a row. In the row from right to left is a frame from the video, the status of the object definitions, and the current calculation of evidence. A key time step is when a new cue is extracted, since otherwise there is no change in the evidence and everything remains the same. In each frame, segmented regions are surrounded with a border and filled with a solid color, activity cues are labeled with all letters capitalized next to a box surrounding the face of the person doing the action, speech cues are white with quotes and object labels are white with all letters capitalized. In each object definition box, each object is in bold followed by the cues associated with it: the type of cue (in italic), the cue value, and the weight. The evidence column shows the current calculation of evidence that a particular region, r_n, is a particular object, o_i, for each region and each object. This leads to an initial mislabeling of the whiteboard (see figure 6(b)) as a projector screen. At this point, most visual based recognition systems would have no recourse to correct the mislabel. MCOR, however, is able to overcome the obstacle by taking into account evidence from other types of cues (see figure 6(c-d)), in this case, pointing, writing, and erasing activity cues.

Thus, more robust object recognition can be produced when the ability to include cues of various types is provided, for this allows the system to be less dependent on the weaknesses of any single type.

6.4 Same Cue Associated with Different Objects

In the second task, two types of objects were to be recognized: laptops and chairs. An additional difficulty was added by having both the laptop and chair objects associated with the same activity, i.e. sitting. Figure 7(d) shows the results of the object recognition task, where despite the difficulty, the two laptops in the scene and a number of chairs were correctly labeled (Some chairs were missed because they were neither sat upon nor matched with the visual cue. Thus, additional cues would have to be learned or provided to recognize the rest).

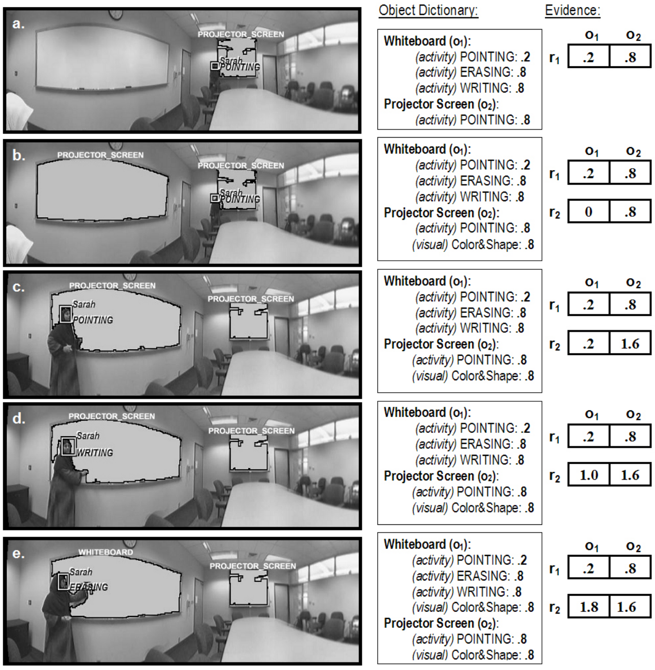


Figure 6: The use of evidence to correct a mislabeled object due to an unreliable visual cue.

Although initially both objects were labeled chair (see figure 7(c)), due to the only cue provided thus far being associated with both objects, this ambiguity was able to be resolved by the algorithm's ability to take evidence from multiple cues, (in this case, additional evidence was given by a speech cue: see figure 7(d)) which aided in distinguishing the laptops from the chairs.

Thus, having the same cue belong to multiple object definitions can lead to ambiguity in the recognition. This is, however, a realistic representation of the world, where it is very likely that the same cue may be correctly ascribed to more than one object, i.e. the set of cues belonging to each object will never be completely disjunct in real world scenarios, and so an algorithm must be able to handle such ambiguity. MCOR is able to do so, as demonstrated, by collecting evidence from multiple cues to decipher the difference.

6.5 Misleading Cue

In the third task, the table was to be recognized, although a chair definition was also provided. For added difficulty, a misleading cue was given. Although in most cases humans interact with objects according to their definition, on occasion, a human may misuse an object or interact act with it in an unconventional manner producing a misleading cue.

In this case, the misleading cue was produced by having the human sit on the table, an activity clearly attached to the function and definition of a chair, but somewhat adverse to that of a table. Initially, the algorithm is indeed confused

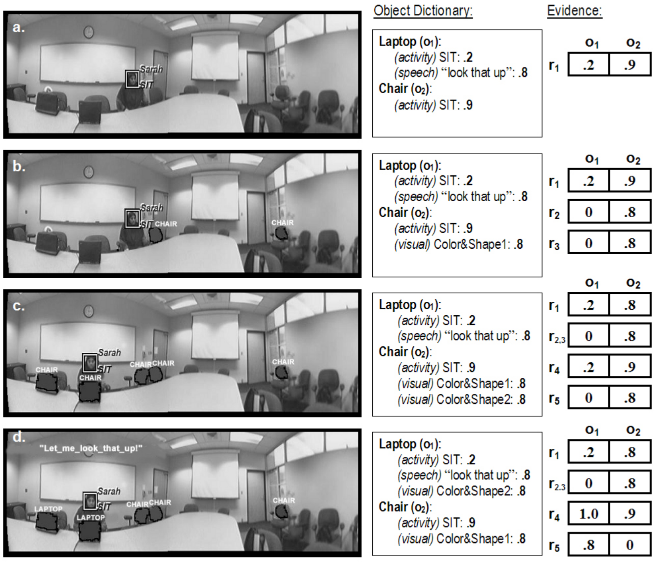


Figure 7: The use of evidence to disambiguate a cue that provides evidence for more than one object.

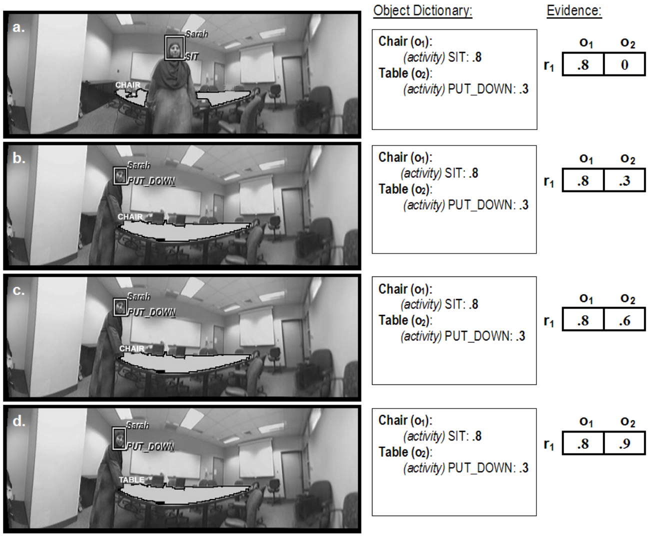


Figure 8: The correction of a false positive caused by a misleading cue.

and incorrectly labels the table, a chair (see figure 8(a)). Although sitting is a strong indication of a chair (as shown by the high value of its weight: figure 8). The algorithm is able to take advantage of the fact that people tend to do the correct action far more frequently than the incorrect, as seen by the putting-objects-down activities in figure 8(b-d), to correct the mistake caused by the misleading cue (see figure 8(d)).

In addition, figure 8 shows how the color and shape of the table which was originally and incorrectly associated with the chair was placed in the proper definition after the correction of the label. A similar case can be noted in figure 7.

Thus, the algorithm was able to demonstrate its robustness to misleading cues through the multiple cue framework.

7. CONCLUSION AND FUTURE WORK

With this paper, we have shown that the numerous cues humans provide when interacting with objects can be taken advantage of to deal with the weakness of particular cue types such as visual cues, the ambiguity caused by cues that may provide evidence for more than one object, and the misleading evidence of cues produced by unconventional use of the object. In other words, more robust object recognition were attained through the multiple cue framework described in this paper.

Work in the future includes developing a framework that not only learns a general object description, but includes context specific ones as well, so that given a specific environment, information useful for recognizing an object, such as the color of a chair, that may be only applicable in that environment can be utilized, where as if only a general description of an object were learned and used, that information would not be taken advantage of and may lead to a more difficult time in recognizing. For example, if in a particular meeting room, all the chairs were red, the algorithm would learn this trait, and so put a lot more reliance on the color of the chair, where as we know in general chairs are not any particular color and so the general description of a chair will not put much weight on color. Thus, if only the general description were used, the useful information of color will not be utilized in a context where it could have been helpful. Future work will attempt to alleviate that problem.

In addition, further advantage of the Probabilistic Relational Model framework will be taken so that weight values can depend not only on the individual cue values of a cue and object label, but on other properties such as the type of cue, the number of the cue, as well as other relationships besides the object-cue one such as the influence of other cues on cues. It can also be adjusted to relate to the context problem described above.

Further goals include applying this framework onto a mobile platform in order to more clearly demonstrate its use on robotic agents.

Other less significant improvements include a more sophisticated segmentation method, since those regions are supposed to represent entire objects and some objects cannot be segmented by a simple color-based region growing scheme. Also, although this is not directly related to the research problem explored, the inclusion of more automated and better cue recognizers (i.e. activity recognizer, speech recognizer), which will reduce the effort needed to label such things manually and demonstrate the use of this framework in integrating cues generated by already exisiting systems.

8. ACKNOWLEDGEMENTS

This research is partly sponsored by SRI International under subcontract no. 03-000211 and by NSF Grant No. IIS-0637069. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies.

9. REFERENCES

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(11):1475–1490, November 2004.

[2] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *IJCAI*, pages 1300–1309, 1999.

[3] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[4] 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.

[5] D. J. Moore, I. A. Essa, and M. H. Hayes. Exploiting human actions and object context for recognition tasks. In *ICCV (1)*, pages 80–86, 1999.

[6] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees: a graphical model realting features, objects, and scenes. *NIPS*, 16, 2003.

[7] E. Murphy-Chutorian and J. Triesch. Shared features for scalable appearance-based object recognition. *Proc. IEEE Workshop Applications of Computer Vision*, January 2005.

[8] S. Palmer, E. Rosch, and P. Chase. Canonical perspective and the perception of objects. In L. J. and A. Baddeley, editors, *Attention and Performance*, pages 135–151. Erlbaum Hillsdale, N.J., 1981.

[9] R. K. Srihari. Computational models for integrating linguistic and visual information: A survey. *Artificial Intelligence Review*, 8(5-6):349–369, 1994.

[10] M. M. Veloso, P. E. Rybski, and F. 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.

[11] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001.

[12] D. Wyatt, M. Philipose, and T. Choudhury. Unsupervised activity recognition using automatically mined common sense. *Proceedings of AAAI-05*, pages 21–27, 2005.