Cross-Modal Learning of Visual Categories using Different Levels of Supervision

Mario Fritz[1], Geert-Jan M. Kruijff[2], and Bernt Schiele[1]

[1] Comuter Science Department, TU-Darmstadt, Germany
{schiele,fritz}@informatik.tu-darmstadt.de

[2] Language Technology Lab, DFKI GmbH, Saarbrücken, Germany
gj@dfki.de

Abstract. Today's object categorization methods use either supervised or unsupervised training methods. While supervised methods tend to produce more accurate results, unsupervised methods are highly attractive due to their potential to use far more and unlabeled training data. This paper proposes a novel method that uses unsupervised training to obtain visual groupings of objects and a cross-modal learning scheme to overcome inherent limitations of purely unsupervised training. The method uses a unified and scale-invariant object representation that allows to handle labeled as well as unlabeled information in a coherent way. One of the potential settings is to learn object category models from many unlabeled observations and a few dialogue interactions that can be ambiguous or even erroneous. First experiments demonstrate the ability of the system to learn meaningful generalizations across objects already from a few dialogue interactions.

Key words: object categorization, cross-modal learning, incremental and interactive learning

1 Introduction

In computer vision, impressive progress has been made recently not only for object identification but also for object categorization in real-world scenes. Quite interestingly, these methods use different learning methods ranging from supervised methods [13], over weakly supervised methods [5] to unsupervised methods [18, 8], and also ranging from generative to discriminant learning methods [6]. Following common practice today these systems are evaluated on predefined training and test sets enabling direct comparisons. For a cognitive vision system however it is highly important that models and representations are flexible and evolvable over time enabling continuous or even life-long learning. This goal is not only much harder to achieve but it is also more difficult to evaluate and compare and consequently it is not clear how the above mentioned approaches could be extended to deal with this more challenging scenario.

As we understand cognitive vision systems one of their most important and fundamental abilities is to evolve over time by actively and passively acquiring

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

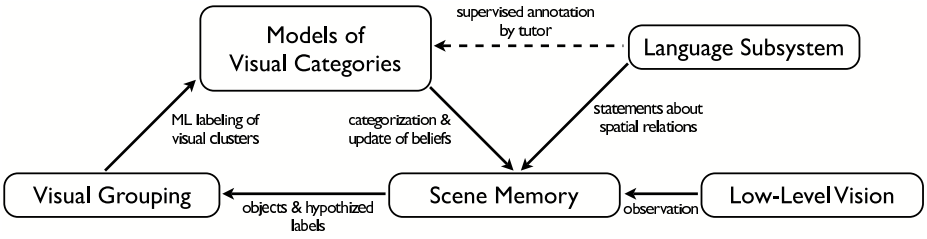


Fig. 1. System overview.

new knowledge and incorporating that knowledge into the system. While there exists a wide range of sources of knowledge, in this paper we focus on the ability to acquire new knowledge through dialogue interactions with humans. In this scenario we can identify a number of requirements a cognitive vision system needs to fulfill. First, to enable interactivity, the representations and models of the systems need to enable incremental processing and learning. Secondly and in order to test and evaluate such systems all processing should be done in real-time or at least at speeds that allow real interactivity. Third, as humans will use language to interact with the system, the learning mechanisms have to allow cross-modal learning from vision and language. And forth the learning algorithms should enable to deal with ambiguous and even erroneous input both from vision and language.

In the following, we present an approach for cross-modal learning of visual categories which integrates language and vision input. Language provides "scene descriptions", describing objects and their spatial relations in a given scene, which provides a top-down description which is then related to the bottom-up generalizations of the vision system. The scene descriptions can be interpreted on ontologically rich knowledge representations, which make it possible to use ontologies to mediate between linguistically expressed meaning, and the categories formed in the visual system. Using the hierarchical structure of ontologies, and the possibility to perform ontological inference over instances on these ontologies, provides a more general and better scalable approach to "visual grounding" of language than provided by the string-based approach proposed in [17], or earlier ontology-based approaches such as [10].

System overview. Figure 1 shows an overview of the presented system, which is tighly related to the structure of this paper. Section 2 describes the vision system which is decomposed into low-level functionality (feature extraction, object discovery and object representation (Sec. 2.1)), the unsupervised visual grouping step (Sec. 2.2) and the categorization procedure (Sec. 2.3). Section 3 describes the language system that parses an utterance to a logical form. Section 4 explains the spatial reasoning processes that associate the expressions with the visual observation which are then used to probabilistically associate labels to the clusters obtained from visual grouping. Section 5 illustrates the functions of the integrated system and provides empirical evidence for our claims.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

2 Vision Sub-System

The description of the vision sub-system is divided into three parts. First, the low-level functionalities are described, that extract local image features, discover object centers and extract a scale-invariant object representation at the hypothesized object positions. Second, the visual grouping procedure is explained, that provides the system with a data-driven generalization over category instances that is obtained in a totally unsupervised manner. This is implemented by an agglomerative clustering on the extracted scale-invariant object representations. Third, the model for performing categorization is described that is based on the same scale-invariant representation and can handle information obtained in a supervised, semi-supervised and unsupervised fashion.

2.1 Low-Level Vision and Object Representation

Feature extraction. When a new image is grabbed from the camera, SIFT descriptors [14] are extracted at hessian-laplace interest points [16]. While there exists a wide range of interest point and descriptor combinations, we decided for this particular combination based on evaluations on different categorization tasks [15]. Following a common philosophy in the field to visual categorization [4, 13, 18, 12, 1, 7], we first generate a visual codebook based on clustering of detected features. In this paper we use a codebook with 1000 entries obtained by k-means clustering. The matching procedure is accelerated by re-normalizing the codebook entries to a fixed length, which transforms the L_2-norm to a scalar product with an additive constant: $||\boldsymbol{x}, \boldsymbol{y}||^2_{L_2} = (\boldsymbol{x} - \boldsymbol{y})^2 = \underbrace{\boldsymbol{x}^2 + \boldsymbol{y}^2}_{const} - 2\boldsymbol{x}\boldsymbol{y}$. The matching of all image features to all codebook entries can now be computed by a simple matrix multiplication which lends itself to further speed-ups. The introduced errors are found to be negligible.

Object representation. One of the important ingredients of our system is that we use a unified and scale-invariant object representation. This representation is the basis not only for discovering objects in the scene, but also for visual grouping and object categorization. More specifically we adapt the representation of scale-invariant patterns from [7]. This representation can be seen as a extension to the bag-of-words representation [4] by adding two spatial dimensions to each bin [12]. In order to obtain a localized, scale-invariant representation, features are collected within a fixed aperture (see also [1]) and the feature positions (pos_x, pos_y) are normalized with respect to the center of the pattern (c_x, c_y): $(pos_x, pos_y)' = ((pos_x - c_x)/\sigma), (pos_y - c_y)/\sigma)$ where σ is the detected scale of the feature. For efficiency, we store these patterns as sparse vectors Ψ.

Object Discovery. We discover objects in the scene as reoccurring patterns as described in [7]. The method can be seen as an adaptive approach for acquiring a feature statistics of objects in recently observed scenes. This statistic is used to hypothesize object centers given the observed patterns by finding maxima of the computed likelihood function. Instead of taking only the global maxima of this likelihood function, we hypothesize all local maxima as object centers. As

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

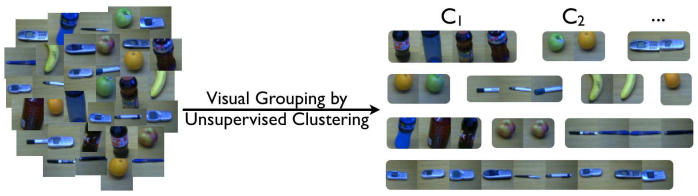


Fig. 2. Visual grouping of objects by clustering.

a result, an observed scene is represented by a set of localized, scale-invariant patterns Ψ at the locations indicated by this discovery procedure.

By keeping a fixed number of patterns (the most recent ones) in memory, this online method for object discovery meets our real-time requirements and also consumes a constant amount of memory. Even though the original method can handle significant background clutter we cannot benefit from that large statistics, as the number of interactions with the system is limited. In a sense we decided to trade generality for real-time capability of the system. As it has been shown that arbitrary backgrounds can be handled when sufficient statistics are available, we will extend our system in this direction.

2.2 Unsupervised Visual Grouping

Similarily to [8], we use an agglomerative clustering scheme (average linkage) to group object instances in an unsupervised manner. The objects are represented as scale-invariant patterns Ψ (Sec. 2.1), which are normalized to unit length and we use the scalar-product to measure similarity between the objects. The threshold required for the clustering scheme was set empirically to a constant value for all our experiments. Figure 2 visualizes the clusters C_1 to C_N obtained by our system given the observed objects displayed on the left. Although there are some confusions, we observe a good generalization across category instances. In order to obtain representatives $\boldsymbol{\Psi}_{C_l}$ for each cluster C_l, we compute a weighted sum of the observed patterns Ψ_k:

$$\boldsymbol{\Psi}_{C_l} = \sum_k p(C_l|\Psi_k)\Psi_k \tag{1}$$

In our implementation, we have chosen to use hard assignment of the patterns to the clusters which renders the probability $p(C_l|\Psi_k)$ of assigning pattern Ψ_k to cluster C_l binary.

2.3 A Joint Model for Visual Categorization from Supervised, Semi-Supervised and Unsupervised Input

In this section, we present a model for visual category recognition that combines different levels of supervision to a joint model. The key ingredient is the scale-invariant pattern representation from Secion 2.1, which we use throughout. The last section (2.2) formulated an unsupervised grouping process using this common representation.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

Supervised Categorization. To provide basic functionality for our system, we describe how supervised categorization can be implemented. Similar to clustering in Section 2.2, we model each category A_i by a single representative $\boldsymbol{\Psi}^S_{A_i}$ (superscript S denotes the supervised model). This is done by summing over all training patterns available for that category

$$\boldsymbol{\Psi}^S_{A_i} = \sum_{j \in \mathbb{S}^{A_i}} \Psi_j, \tag{2}$$

where $\mathbb{S}^{A_i}$ denotes the indices of the patterns that are labeled with category A_i in a supervised manner (e.g. "This is a bottle").

Incorporating Semi-Supervision and Unsupervised Information. We formulate the fusion of information obtained from supervised to unsupervised sources as an extension of the supervised case by assuming uncertainty about the correct labeling of the clusters C_l and their representatives $\boldsymbol{\Psi}_{C_l}$ from the unsupervised visual grouping step (Sec. 2.2):

$$\boldsymbol{\Psi}_{A_i} = \underbrace{\boldsymbol{\Psi}^S_{A_i}}_{\text{supervised}} + \underbrace{\sum_l p(A_i|C_l) \overbrace{\boldsymbol{\Psi}_{C_l}}^{unsupervised}}_{\text{semi-supervised}} \tag{3}$$

$p(A_i|C_j)$ encodes the belief that cluster C_l contains instances of category A_i. How this probability is computed from a few interactions and updated by associating spatial expression with visual observations is described in Section 4.

To perform classification in the supervised and semi-supervised case, we evaluate the proposed model $\boldsymbol{\Psi}_{A_i}$ as well as $\boldsymbol{\Psi}^S_{A_i}$ for an observed pattern $\bar{\Psi}$ by using histogram intersection. Intuitively, the intersection measures to which percentage the model explains the observation, which we interpret as probability of belonging to the same class. In order to make models and observations comparable we normalize both to one. Bayes' rule is applied afterwards to obtain the model posterior:

$$p(\bar{\Psi}|\boldsymbol{\Psi}_{A_i}) = \sum \min(\bar{\Psi}, \boldsymbol{\Psi}_{A_i}) \tag{4}$$

$$p(\boldsymbol{\Psi}_{A_i}|\bar{\Psi}) = \frac{p(\bar{\Psi}|\boldsymbol{\Psi}_{A_i})p(\boldsymbol{\Psi}_{A_i})}{\sum_A p(\bar{\Psi}|\boldsymbol{\Psi}_{A_i})p(\boldsymbol{\Psi}_{A_i})} \tag{5}$$

$p(\boldsymbol{\Psi}_{A_i})$ is the category prior, which we assume to be uniform. We decide for the category label with the highest posterior:

$$\hat{A}_i = \underset{A_i}{\operatorname{argmax}}\, p(\boldsymbol{\Psi}_{A_i}|\bar{\Psi}) \tag{6}$$

3 Language Sub-System

In human-assisted visual learning, a human tutor provides the system with descriptions of the current visual scene. To relate these descriptions to the visual

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

input, the system constructs a representation of the meaning of an utterance. For this analysis we use a Combinatory Categorial Grammar[3] parser[1]. The parser uses a CCG grammar to relate the syntactic structure of an utterance to the propositional meaning it expresses. Meaning is represented as an ontologically richly sorted, relational structure similar to a description logic formula [2], which makes it possible to use ontologies to mediate between linguistically expressed meaning, and the categories formed in the visual system. Using the hierarchical structure of ontologies, and the possibility to perform ontological inference over instances on these ontologies, provides a more general and better scalable approach to "visual grounding" of language than provided by the string-based approach proposed in e.g. [17], or previous ontology-based approaches such as [10, 11].

In our scenario, utterances are typically predicative copulative sentences in indicative mood (i.e. "X is Y"), which assert that a given predication ("Y") holds for the subject of the sentence ("X"). In our examples, the predication consists of a phrase that encodes a spatial relation (e.g. "left of the bottle" or "below the apple"). In the logical form, the subject is represented as the <Restr> of the state description that is denoted by the utterance, whereas the predication is represented as <Scope>.

We can thus easily derive the spatial configuration asserted in an utterance from its logical form representation. The following example shows such a logical form that is the result of the parsing process of the utterance "the mobile is left of the bottle":

```
@b1:state(be ^
          <Mood>ind ^
          <Restr>(m1:thing ^ mobile ^
                  <Delimitation>unique ^
                  <Number>sg ^
                  <Quantification>specific_singular) ^
          <Scope>(l1:region ^ left ^
                  <Plane>horizontal ^
                  <Positioning>static ^
                  <Dir:Anchor>(b2:thing ^ bottle ^
                               <Delimitation>unique ^
                               <Number>sg ^
                               <Owner-of>+ ^
                               <Quantification>specific_singular)))
```

For entering utterances into the system, we connect the parser with a speech recognizer as well as a keyboard interface. In the experiments, we mostly use the keyboard interface as well as scripted input for larger evaluations.

4 Spatial Reasoning and Cross-Modal Association

Modeling spatial relations as perceived by the human is a challenge in itself, as issues like reference frame and context have to be handled appropriately in situated dialogue systems [9]. Considering the scenarios and main focus of this paper, we restricted ourselves to modeling four basic spatial relations $R \in$

[1] `http://openccg.sourceforge.net`

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

{"leftof","rightof","above","below"}. We employ triangular shaped distributions $p(\text{pos}(\Psi_i), \text{pos}(\Psi_j)|R)$ defined in 2d image coordinates, where objects are referenced by their patterns Ψ_i and $\text{pos}(\Psi_i)$ denotes their position in image coordinates. Although these distributions are represented as non-parametric kernel densities which lend themselves to online updating, we don't explore this option in this paper and keep them fixed in the experiments.

Spatial Reasoning. We formulate the association of a spatial expression E extracted from an utterance (see Sec. 3) with two patterns Ψ_i and Ψ_j with positions $\text{pos}(\Psi_i)$ and $\text{pos}(\Psi_j)$ observed in scene S_k, as finding the most likely pair $\hat{P}_{i,j}$ of patterns: $\hat{P}_{i,j}^{(k)} = \text{argmax}_{P_{i,j}}\, p(P_{i,j}|E, S_k)$,where

$$\begin{aligned} p(P_{i,j}|E, S_k) &= p(\Psi_i, \Psi_j, \text{pos}(\Psi_i), \text{pos}(\Psi_j)|E, S_k) \\ &= p(\Psi_i|E, S_k)\; p(\Psi_j|E, S_k)\; p(\text{pos}(\Psi_i), \text{pos}(\Psi_j)|E, S_k), \end{aligned} \tag{7}$$

with

$$p(\Psi|E, S_k) = \sum_h p(\Psi|A_h) p(A_h|E, S_k). \tag{8}$$

As we don't model a complete category system yet, leave out contextual effects and assume certainty about the expression E referring to the categories A_{e_1} and A_{e_2} and the relation R, the equation simplifies to

$$p(P_{i,j}|E, S_k) = p(\Psi_i|A_{e_1}) p(\Psi_j|A_{e_2}) p(\text{pos}(\Psi_i), \text{pos}(\Psi_j)|R) \tag{9}$$

Finally, we insert the visual model from Eq. 3 to obtain a computational model:

$$p(P_{i,j}|E, S_k) = p(\Psi_i|\boldsymbol{\Psi}_{e_1}) p(\Psi_j|\boldsymbol{\Psi}_{e_2}) p(\text{pos}(\Psi_i), \text{pos}(\Psi_j)|R) \tag{10}$$

This formulation facilitates incorporating information and belief from previous interactions as well as learning from scratch. If no information about the visual categories is available $p(\Psi_i|\boldsymbol{\Psi}_{e_1})$ and $p(\Psi_j|\boldsymbol{\Psi}_{e_2})$ become uninformative and the system relies only on its notion of spatial relations $p(\text{pos}(\Psi_i), \text{pos}(\Psi_j)|R)$. This can lead to wrong associations. In Section 5 we present an example and show that the system can successfully deal with this issue.

Cluster Labeling. We want to make use of the belief about associations between spatial expressions (Eq. 7) and objects in the scene to improve the label assignment $p(A_i|C_l)$ of the object clusters in Eq. 3. Therefore we accumulate the evidence for cluster C_l being labeled as containing instances of category A_i by a simple count statistic $p(C_l|A_i)$ based on the maximum likelihood estimates of Equation 7. The probability for assigning label A_i to cluster C_l is obtained by applying Bayes' rule

$$p(A_i|C_l) = \frac{p(C_l|A_i)p(A_i)}{\sum_i p(C_l|A_i)p(A_i)}, \tag{11}$$

where we assume a uniform category prior $p(A_i)$. This closes the loop in our system as outlined in Fig. 1.

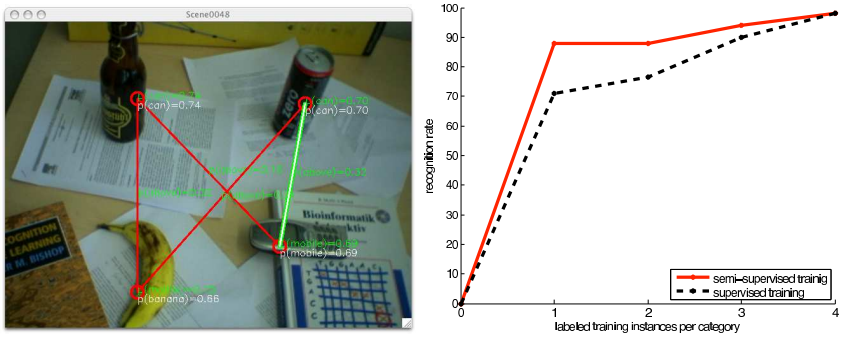
Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

5 Experiments

In the first part of our experiments, we describe two scenarios, that show the capabilities of our system to propagate information, resolve ambiguities and recover from errors. In the second part we perform a quantitative analysis to show that the unsupervised visual grouping step improves learning speed and accuracy with respect to the amount of provided supervision. Finally, we'll provide computation times for the individual modules to judge about the real-time capabilities of the system.

5.1 Label Propagation and Conflict Resolution

Scenario 1 - Label forward propagation. In the first scenario, one annotated example each for banana and mobile is presented to the system. Then the system observes the scene as shown in the screenshot in Figure 3(a) and the utterance "the can is above the mobile" is parsed. The red lines visualize the observed relations between objects in the scene. Very unlikely ones have already been pruned away by the system. By generalizing across category instances, system identifies "mobile" and "banana" correctly (with probabilities 0.69 and 0.66 respectively) while evaluating "mobile" model for the banana results in a low probability of 0.15. Consequently the most likely relation is inferred correctly and displayed in light green. A model for the category "can" is created and the observed mobile is added to the existing model for "mobile". In fact, the figure shows the state in which the acquired "can" model is already used for detection. The can is detected correctly, but also the bottle gets a high score for the "can" model, as it's the best explanation given the learned categories (banana, mobile, can).



(a) (b)

Fig. 3. (a) shows an example scenario for propagation of labels from known categories to unknown ones. (b) shows the improvement of the semi-supervised approach over the purely supervised approach by exploiting information from unlabeled data.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

Scenario 2 - Label backward propagation. In the second scenario, we show how the system can recover from erroneous beliefs and update its models accordingly. The system starts without any knowledge about visual categories. Figure 4 shows a screenshot displaying the scene as observed by the system, which is accompanied by the utterance "the can is above the mobile". Using the same visualization as in the previous scenario, it can be seen in the left image that the most likely relation inferred by the system is wrong. Now we provide the system with supervised knowledge of the visual categories bottle and pen. Revisiting the scene in memory, the object probabilities get updated and the belief about the associated relation gets changed to the correct one as shown in the right image.

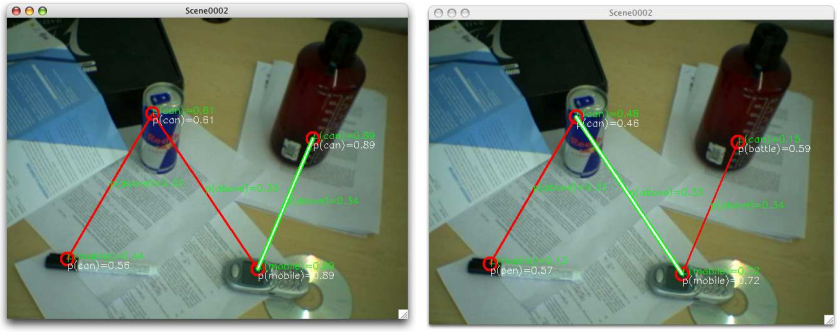


Fig. 4. Scenario which shows how the system updates associations (light green) to recover form an incorrect belief.

5.2 Quantitative Evaluation

Semi-Supervised Learning. Finally, we performed a quantitative analysis by taking 2 images of 5 instances for each of the categories: mobile phone, pen, bottle, can and apple. We gradually increase the training set from one instance for each category to four instances. Figure 3(b) shows that the semi-supervised learning (Sec. 2.3) outperforms the purely supervised learning, as the few available labels get propagated to the unlabeled data (Eq. 11) which was clustered by the visual grouping step (Sec. 2.2). The experiments were performed using 5 fold cross-validation.

Speed. The system as described in this paper runs at about 1Hz on a CoreDuo 2GHz laptop when detection and categorization are performed. An update of the clustering and spatial reasoning takes about 2 seconds total. Therefore the system is fast enough to operate interactively with a human tutor.

6 Conclusions

We present a system for cross-modal learning that combines unsupervised and supervised information in a unified framework. The mechanism that associates

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

expressions from language with the visual input can resolve ambiguous input and recover from erroneous beliefs. The experimental section provides qualitative as well as quantitative results that show these capabilities of the system. Finally, we were able to cut down the computing time to a level at which a human can interact with the system as a tutor.

Acknowledgments: This work has been funded, in part, by the EU project CoSy (IST-2002-004250).

References

1. Ankur Agarwal and Bill Triggs. Hyperfeatures - multilevel local coding for visual recognition. In *ECCV'06*. Springer, 2006.
2. Jason Baldridge and Geert-Jan M. Kruijff. Coupling ccg and hybrid logic dependency semantics. In *ACL '02*, Morristown, NJ, USA, 2001.
3. Jason Baldridge and Geert-Jan M. Kruijff. Multi-modal combinatory categorial grammar. In *EACL '03*, Morristown, NJ, USA, 2003.
4. G. Csurka, C.R. Dance, L. Fan, J. Willarnowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV'04 Workshop on Stat. Learn. in Comp. Vis.*, pages 59–74, Prague, Czech Republic, May 2004.
5. R. Fergus, A. Zisserman, and P. Perona. Object class recognition by unsupervised scale-invariant learning. In *CVPR'03*, 2003.
6. M. Fritz, B. Leibe, B. Caputo, and B. Schiele. Integrating representative and discriminant models for object category detection. In *ICCV'05*, Beijing, China, October 2005.
7. M. Fritz and B. Schiele. Towards unsupervised discovery of visual categories. In *DAGM'06*, Berlin, Germany, September 2006.
8. K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In *CVPR'06*, pages 19–25, Washington, DC, USA, 2006. IEEE Computer Society.
9. John Kelleher, Geert-Jan Kruijff, and Fintan Costello. Proximity in context: an empirically grounded computational model of proximity for processing topological spatial expression. In *Coling-ACL '06*, Sydney Australia, 2006.
10. Geert-Jan M. Kruijff, John D. Kelleher, Gregor Berginc, and Aleš Leonardis. Structural descriptions in Human-Assisted robot visual learning. In *Proceedings of 1st Annual Conference on Human-Robot Interaction*, March 2006.
11. Geert-Jan M. Kruijff, John D. Kelleher, and Nick Hawes. Information fusion for visual reference resolution in dynamic situated dialogue. In *PIT 2006*, Kloster Irsee, Germany, June 2006.
12. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR'06*, pages 2169–2178, Washington, DC, USA, 2006.
13. B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR'05*, San Diego, CA, USA, June 2005.
14. D. Lowe. Object recognition from local scale invariant features. In *ICCV'99*, 1999.
15. K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. In *ICCV'05*, Beijing, China, October 2005.
16. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *CVPR'03*, 2003.
17. D.K. Roy. Learning words and syntax for a scene description task. *Computer Speech and Language*, 16(3), 2002.
18. J. Sivic, B. C. Russell, A. A. Efros, Andrew Zisserman, and William T. Freeman. Discovering objects and their locations in images. In *ICCV'05*, Beijing, China, October 2005.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de