Learning a Knowledge Base of Ontological Concepts for High-Level Scene Interpretation

Johannes Hartz
Cognitive Systems Laboratory, Department of Informatics
Hamburg University
hartz[at]informatik.uni-hamburg.de

Bernd Neumann
Cognitive Systems Laboratory, Department of Informatics
Hamburg University
neumann[at]informatik.uni-hamburg.de

***Abstract*— *Ontological concept descriptions of scene objects and aggregates play an essential role in model-based scene interpretation. An aggregate specifies a set of objects with certain properties and relations which together constitute a meaningful scene entity. In this paper we show how ontological concept descriptions for spatially related objects and aggregates can be learnt from positive and negative examples. Our approach, based on Version Space Learning introduced by Mitchell ([1]), features a rich representation language encompassing quantitative and qualitative attributes and relations. Using examples from the buildings domain, we show that aggregate concepts for window arrays, balconies and other structures can in fact be learnt from annotated images and successfully employed in the conceptual knowledge base of a scene interpretation system. Furthermore we argue that our approach can be extended to cover ontological concepts of any kind, with very few restrictions.*

I. Introduction

In computer vision, growing interest in artificial cognitive systems has brought about increased efforts to extend vision systems towards capabilities for high-level vision or scene interpretation. These are terms commonly used for vision tasks going beyond single-object recognition, such as inferring the existence and location of occluded aggregate parts from already observed ones. As explicated in [4], scene interpretation can be modelled formally as a knowledge-based process. The burden of the interpretation process lies on the conceptual descriptions, and the richer a domain, the more demanding is the task of designing these descriptions. It is foreseeable that designing knowledge bases for larger applications using a handcrafting approach will be prohibitively costly and error-prone.

We therefore started to investigate supervised learning in the eTRIMS project, with the belief that in the long run high-level vision can only be achieved by leading the system through a supervised learning phase where the concepts for a particular domain are acquired based on examples. Different from a probabilistic approach (e.g. [5], [6], [7]), we chose the representation language used in our scene interpretation system SCENIC [8], [23] which represents variability in terms of ranges with crisp boundaries or enumeration of possible values. Apart of the fact that this way we can evaluate the learnt concepts by applying them to real-world scenes through the SCENIC system, this approach also allows us to invoke and extend well-known learning methods from symbolic AI. Our approach is in the spirit of the seminal work of Winston [19] who showed how spatial structures in the blocks-world could be learnt. We rephrase this problem for a more general domain by using the Version Space Learning framework.

Our main contributions are

- developing a description language for spatial object arrangements,
- applying the learning procedure to a concrete real-world domain, and
- evaluating the results in an operational scene interpretation system.

In the next section we present a strong motivation to choose Mitchell's Version Space Learning framework for concept learning aimed at machine interpretation. Then we present the concept language, which is designed to allow realistic concept descriptions. Section III deals with the problem of hypothesis selection which arises when several concept descriptions correctly cover all positive and negative examples. This is the rule rather than the exception in Version Space Learning. In Section IV we argue that a comprehensive knowledge base of ontological concepts of any kind can be learnt in a Version Space framework, given general-specific orderable concept attributes and a finite set of concept relations. We also introduce the term *concept differentiation* as a quality measure for a conceptual knowledge base and present an approach for learning maximally diverse concepts. In Section V we present experimental results for the application domain of building facades. Section VI, finally, presents conclusions and an outlook on further work.

Fig. 1. Annotated training image with four instances of aggregate "Entrance"

This research has been supported by the European Community under the grant IST 027113, eTRIMS - eTraining for Interpreting Images of Man-Made Scenes

II. Version Space Learning of ontological concepts

A. Inductive bias and interpretation properties

For every form of Inductive Learning a resulting concept hypothesis h has to approximate the correct output classification $h(e) \leftrightarrow c(e)$, even for examples that have not been shown during training. This property is commonly referred to as the $Inductive\ Leap$. Without any additional assumptions, this task cannot be solved [3], so the need of an inductive bias is at the heart of any Inductive Learning process. In fact, the inductive bias is the necessary assumption which possible target concept to prefer over another (a classic example is Occam's Razor / MDL). But despite the learning process' need of an inductive bias, we also have to consider our general aim of concept learning for machine interpretation. For the generic interpretation process to be able to perform as flexible as possible, we do not want to bias the concept learning process beyond the intrinsic need. Version-Space Learning lends itself to this task perfectly. The Version-Space learning process is bias-free, because all hypotheses consistent with the training data are induced (II-B). An inductive bias is introduced only through the concept language, not the learning procedure. If the concept language allows any subsets of training instances $I_1 \subset I$ and $I_2 \subset I$ to be generalized to the same hypothesis h, than the language is biased. By employing Version Space Learning we have full control over this inductive bias. Our concept language is presented in section II-C. Section III shows how we can exploit the fact that Version Space Learning is bias-free additionally to derive a **confidence criterion** for later generic machine interpretation.

B. Learning Procedure

Version Space Learning ([1], [2]) is a framework for supervised concept learning, i.e. learning by means of positive and negative examples given by a teacher. During the learning process, the space of possible concept hypotheses VS is implicitly represented through an upper and a lower bound on the generality of the hypotheses $h \in VS$. The General Boundary GB contains all maximally general members of VS, the Specific Boundary SB contains all maximally specific members of VS. GB and SB completely determine VS as the set of hypotheses h being $more\text{-}general\text{-}or\text{-}equal$ to an element of SB and $more\text{-}specific\text{-}or\text{-}equal$ to an element of GB.

Initially, GB includes all possible training examples and SB excludes all possible training examples. As a positive example e^+ is presented, SB has to be generalised to include e^+. As a negative example e^- is presented, GB must be specialised to exclude e^-. Both, generalisation and specialisation steps, are chosen to be minimal in the sense that as few instances as possible besides e^+ or e^- are included or excluded, respectively. In contrast to minimal generalisations, the minimal specialisation of a hypothesis h leads to a set of hypotheses $\{h', h'', ...\}$, at least for non-trivial cases. For the sake of compactness, more elaborated representation schemes ([9], [10], [11]) and training procedures (([12], [13]) for Version Space Learning are omitted here. Theoretic considerations for inductive concept learning can be found in [14].

The representation of the Version Space by the two boundaries, together with an appropriate revision strategy, allows every hypothesis that is consistent with the training data to be generated. Note, that what is required in order to realise this type of representation is a concept language that constitutes concepts which satisfy the properties of a concept lattice ([15]), i.e. a partial general-specific ordering can be imposed on them.

For the application to the eTRIMS domain, annotated images of building facades are used as input for the learning process. In these images meaningful scene objects have been segmented and labeled (Fig. 1). Scene aggregates are specified solely through a set description of their parts.

For the actual training process we use a concept language with attribute types presented in the next section. To give an intuition of the nature of the concept language beforehand, we present an abbreviated concept description for the aggregate "Entrance", generalised from the four positive examples in Fig. 1:

Size and configuration	
Aggregate Width =	[184..216] cm
Aggregate Height =	[299..366] cm
Composition	
Has-Parts =	[3..4]
door =	[1..1]
stairs =	[1..1]
canopy =	[0..1]
railing =	[0..1]
sign =	[0..1]
Symbolic attributes	
Shape =	{ Quadratic }
Internal spatial relations	
(stairs011) BelowNeighbourOf [0..2] cm (door012)	
(door012) AboveNeighbourOf [0..2] cm (stairs011)	

TABLE I
GENERALISED AGGREGATE DESCRIPTION "ENTRANCE"

C. Representation

In this section we describe the attribute types used to formulate concept descriptions. We also specify generalisation and specialisation criteria which can be used to determine their general-specific ordering (denoted $\leq$ and $\geq$). A methodology to compute minimal concept attribute generalisations (denoted $\uparrow$) is presented, which is needed to extend attributes of concept hypotheses $h \in SB$ to cover attribute values of positive examples e_i^+. Specialisation methods (denoted $\downarrow$) to exclude attribute values of negative examples e_i^- from attributes in concept hypotheses $h \in GB$ are also presented.

1) Symbol set type: The symbol set type describes disjunctive symbolic or discrete numerical attribute values.

Example: $Colour = \{Red, Green, Blue\}$

- General-specific ordering of symbol sets S_1 and S_2 of disjoint symbols $\{s_1, s_2, ...\}$:
  - Iff $S_1 \supseteq S_2$: $S_1 \geq S_2$
- Obtaining symbol set S_3 from $S_1 \in h \uparrow S_2 \in e^+$: $S_3 = S_1 \cup S_2$
- Obtaining symbol sets S_i from $S_1 \in h \downarrow S_2 \in e^-$: $\forall s_i \in S_1 \wedge s_i \in S_2 : S_i = S_1 \setminus \{s_i\}$

2) Range type: The range type describes a convex range of metric attribute values. Specialized ranges can have (half) open boundaries due to the exclusion of discrete values. Ranges can contain symbolic infinity values -INF and INF.

Example: $Aggregate\ Height = [160..\text{INF}]$

- General-specific ordering for ranges $R_1 = [l_1..u_1]$ and $R_2 = [l_2..u_2]$:
  - Iff $R_1 \supseteq R_2$: $R_1 \geq R_2$
- Obtaining range R_3 from $R_1 \in h \uparrow R_2 \in e^+$:
  - Iff $l_2 < l_1$: $R_3 = [l_2..u_1]$
  - Iff $u_2 > u_1$: $R_3 = [l_1..u_2]$
  - Iff $l_2 < l_1 \wedge u_2 > u_1$: $R_3 = [l_2..u_2]$
- Obtaining range R_3 from $R_1 \in h \downarrow R_2 \in e^-$:
  - Iff $u_1 > u_2 \wedge l_1 \leq u_2$: $R_3 =]u_2..u_1]$
  - Iff $l_1 < l_2 \wedge u_1 \geq l_2$: $R_3 = [l_1..l_2[$

Note that theoretically the two cases of specialisation may be applicable at the same time, but because the Version Space enforces all $h_i \in GB$ to be more general than any $h_i \in SB$, at most one of the two possible specialisations is valid.

3) Composition type: The composition type describes
1. the number of aggregate parts by the range attribute N and
2. the different part types by the symbol set attribute TN and
3. the number of parts of each type by the subordinate range attributes $T_{1..n}$ in TN.

$$N = [MAX_{i \in n}(l_i) \leq nl \leq \Sigma_{i \in n}(l_i)..\ MIN_{i \in n}\ (u_i) \leq nu \leq \Sigma_{i \in n}(u_i)]^1$$

$$TN = \{T_1 = [l_1..u_1],\ T_2 = [l_2..u_2],...,\ T_n = [l_n..u_n]\}$$

$$\text{Example: } Has-Parts = [3..6]$$
$$Triangle = [2..3]$$
$$Square = [1..3]$$

- General-specific ordering for compositions C_1 and C_2:
  - Iff $N_1 \geq N_2 \wedge \forall T_i \in TN_2 \geq T_i \in TN_1$: $C_1 \geq C_2$

To generalise the compositional properties of an aggregate, all ranges in the composition can be treated individually.

- Obtaining composition C_3 from $C_1 \in h \uparrow C_2 \in e^+$:
  - $N_3 = N_1 \uparrow N_2$, $TN_3 = TN_1$, $\forall T_i \in TN_3 = T_i \in TN_1 \uparrow T_i \in TN_2$

When specialising the composition we have to consider dependencies between the total number of parts and the number of parts per type explicitly.

- Obtaining C_3 from $N_1 \in h \downarrow N_2 \in e^-$ might lead to the same specialisation step for subranges T_i in set TN_3:
  - Iff $nu_1 > nu_2 \wedge nl_1 \leq nu_2$: $N_3 =]nu_2..nu_1]$, $TN_3 = TN_1$
  - Iff $nl_1 < nl_2 \wedge nu_1 \geq nl_2$: $N_3 = [nl_1..nl_2[$, $TN_3 = TN_1$, $\forall T_i \in TN_3$: $T_i = [l_i..\ MIN(u_i, nl_2)[$
- Obtaining C_3 from $T_i \in TN_1 \in h \downarrow T_i \in TN_2 \in e^-$ might lead to a specialisation step of N_3:
  - Iff $u1_i > u2_i \wedge l1_i \leq u2_i$: $N_3 = [MAX(nl_1, \Sigma_{i \in n}(l3_i))..nu_1]$, $TN_3 = TN_1$, $T_i \in TN_3 =]u2_i..u1_i]$
  - Iff $l1_i < l2_i \wedge u1_i \geq l2_i$: $N_3 = [nl_1..MIN(nu_1, \Sigma_{i \in n}(u3_i))]$, $TN_3 = TN_1$, $T_i \in TN_3 = [l1_i..l2_i[$

4) Predicate type: The predicate type represents a freely definable n-ary boolean function over part attribute values. Predicates $p_1..p_n$ are organised in a set P.

Example: $Predicates = \{FuzzyEqual(Parts - Area)\}$

Note that since predicates constrain attribute values, they behave contrarily to symbol sets!

- General-specific ordering for predicates in sets P_1 and P_2:
  - Iff $P_1 \subseteq P_2$: $P_1 \geq P_2$
- Obtaining predicate set P_3 from $P_1 \in h \uparrow P_2 \in e^+$:
  - $P_3 = P_1 \cap P_2$
- Obtaining predicate set P_3 from $P_1 \in h \downarrow P_2 \in e^-$:
  - $\forall p_i \notin P_2 : P_i = P_1 + p_i$

5) Spatial relation type: Spatial relations are learnt between the parts of an aggregate and between the aggregate and possible sourrounding entities, which might be scene objects or other aggregates. To represent the spatial relation between two objects, we employ an 8-neighbourhood to obtain a finite set of possible relations. For this purpose the bounding box of an object induces the eight octants of its neighbourhood: {Left, AboveLeft, Above, AboveRight, Right, BelowRight, Below, BelowLeft}. To quantise spatial relations we use the Euclidean distance d between the related objects' boundaries.

Example: $SR = \{\text{(triangle003) Above } [45..45] \text{ (Square012)}\}^2$

Each spatial relation is a 4-tuple. Spatial relations $l_1..l_n$ are organized in a set L and are treated like predicates.

$l_{i \in n} =$ (object $p1$, relation r, object $p2$, range d)
$L = \{l_1, l_2, .., l_n\}$

In general, for object p_1 and relation type r_j several relations $l_i = (p_1, r_j, p_i, d_i)$ involving different objects p_i may be possible. The relation minimising d_i is called the neighbour relation. Neighbour relations are a specialisation of spatial relations, hence spatial relations form their own general-specific hierarchy. This hierarchy must be considered when performing generalisation and specialisation steps on spatial relations.

[1]The actual values of nl and nu depend on preceding generalisation or specialisation steps

[2]For a textual representation of spatial relations, arbitrary object indices are kept to disambiguate relational structures

- General-specific ordering for spatial relations in sets L_1 and L_2:
  - Iff $L_1 \subseteq L_2 \wedge \forall l_i \in L_1 \geq l_i \in L_2 : L_1 \geq L_2$
- Obtaining spatial relation set L_3 from $L_1 \in h \uparrow L_2 \in e^+$:
  - $L_3 = L_1 \cap L_2, \forall l_i \in L_3 = l_i \in L_1 \uparrow l_i \in L_2$
- Obtaining spatial relation set L_3 from $L_1 \in h \downarrow L_2 \in e^-$:
  - $L_3 = L_1, \forall l_i \in L_3 = l_i \in L_1 \downarrow l_i \in L_2$
  - $\forall l_i \notin L_2 : L_3 = L_1 + l_i, d_i = [0..\text{INF}]$[3]

Note that the particular spatial relation type presented here is just one example of how to impose a symbolic relation. Any other finite set of symbolic relations could be treated accordingly. The spatial relation type includes a range attribute to represent the parts distance. In general, a relation can be enhanced with concept attributes of any type, however the specialization methodology needs to be enhanced then, too.

III. Hypothesis selection

After the learning process has been conducted, the boundary sets SB and GB contain the minimally and the maximally generalised concept hypotheses over all training examples. The space of applicable hypotheses VS covers these two boundary sets and the space in between them. In principle, one could use any member of VS as a classifier. To use the whole VS as a classifier, one could employ a voting scheme. For high-level scene interpretation, however, we are interested in concise concept descriptions which can be included in the conceptual knowledge base of the interpretation system. Therefore we define additional learning objectives by which to select concept hypotheses from a learnt VS.

A. Learning objectives

Assuming a set of positive and negative examples as training data we propose two disjoint learning objectives:

For a given set of training examples we want to learn a concept description

1) which is the most specific representation of training example properties introduced through positive examples. We can assign this concept hypothesis a **high confidence**, because it has strong support from former experience.
2) which is the most general representation of training example properties introduced through positive examples, but still excludes all negative training examples. This concept hypothesis only has weak support from former experience, but could not be proven wrong. Hence it has a **low confidence**.

These learning objectives yield knowledge acquisition which (1) spans the whole space of applicable hypotheses and therefore allows most comprehensive aggregate recognition in the interpretation process, and (2) introduces a confidence measure we can use to evaluate the intepretation result.

[3]The range is opened maximally to satisfy the requirement for minimal specialisation

B. Selection methods

Based on the learning objectives presented above, we now need means to select concept hypothesis from a learnt Version Space VS. Trivially, hypothesis $h \in SB$ satisfies our first learning objective and can be chosen for interpretation purposes. It represents the concept description with the highest confidence h_s.

Every hypothesis $h_i \in GB$ satisfies our second learning objective. If the set GB has converged to one concept hypothesis, this hypothesis is chosen as most general concept description h_g. If GB contains multiple concepts (which is likely considering [16]), we need a selection method to choose a concept hypothesis from GB. But since the hypotheses $h \in GB$ cannot be ordered in a general-specific manner, there is no preference measure to choose a concept hypothesis that can be derived from our learning objective. Several approaches can be considered to overcome this selection problem:

1) A concept hypothesis can be chosen randomly from GB, as proposed in [17], [18].
2) The minimum amount of attribute specialization can be considered to be the selection criterion. This criterion is not applicable for concept languages with a mixture of symbolic and metric attribute types, because these types cannot be compared with regard to the amount of specialisation.
3) The logical conjunction of all $h_i \in GB$ yields a single hypothesis h_g. Hypothesis h_g is the most general concept hypothesis excluding negative examples through all discriminating attributes. Hypothesis h_g is defined for any state of GB, hence it can be chosen as concept hypothesis.

Since every form of hypothesis selection is in fact a form of biasing, we consider the last approach to be the soundest, because it emphasises all attributes that have been used to discriminate negative examples. Approach 1. and 2. (if applicable) lead to an arbitrary selection of discriminating attributes. A refinement of selected concepts is possible for all above approaches via feedback learning as presented in Section V-C.

IV. Building a Knowledge Base of Ontological Concepts for Interpretation

As we have learnt two different concept descriptions for the training set of every ontological entity, we introduce these two descriptions into the knowledge base of our scene interpretation system and relate them taxonomically (h_s is a specialization of h_g). The two concept descriptions give us means to recognize instances of the target concept in every possible extent covered by the positive and negative training examples. Additionally, we can give a confidence value for each recognition. This confidence value is based on the degree of containedness / the distance between the concept description of the recognized instance and h_s and h_g, respectively.

So far, we have considered learning and hypothesis selection with the goal of establishing concept descriptions for individual entities. But for scene interpretation and many other

applications, we want to learn a comprehensive knowledge base with ontological concepts for different real-world entities. Generally, an ontological concept consists of concept attributes and relations to other concepts. Concept attributes can be represented and learnt as shown in II-C - more elaborated attribute types can also be employed, as long as a general-specific ordering can be imposed. Basic relations between ontological concepts are compositional relations and taxonomical relations. The composition of concepts is learnt as presented in II-C.3. Taxonomical relations between concepts can be inferred after learning by applying the general-specific ordering methodology to concepts (Section II-C). Further symbolic relations between concepts can be represented and learnt analogous to the spatial relation type presented in II-C.5. Any symbolic relation can be enriched with further attribute types as mentioned above.

Since our learning approach covers all properties of individual ontological concepts, learning a set of these concepts yields a comprehensive knowledge base. Each concept in this knowledge base has a most specific and a most general representation, each with a set of attributes and relations to other concepts. The knowledge base can be exploited through any form of ontology reasoning, which is typically performed using description logics and a constraint system solver.

For our application to the eTRIMS domain, the learnt concepts in the knowledge base are related compositionally, taxonomically and through spatial relations (Table II, Figure 2). An additional composition attribute is kept to trace transitive compositional relations (e.g. object o_1 is part of o_3 through being part of o_2), which simplifies interpretation. The interpretation process is performed by the SCENIC system.

An important property of ontological concepts is disjointness. To test two concepts for disjointness one can simply construct the logical conjunction of these concepts and check the resulting concept for consistency. If the resulting concept description is inconsistent, the basic concepts are disjoint. We employ disjointness of SB concept hypotheses for our approach to concept differentiation, which is presented in the next section.

A. Concept differentiation

Learnt concept descriptions must be evaluated with respect to existing concepts. Intuitively, we want to make sure that the conceptual descriptions in the knowledge base do not only reflect arbitrarily chosen positive and negative examples but are also constructed to differentiate between each other. We call this quality criterion *concept differentiation*.

Fortunately, the concept learning process can be controlled to yield a knowledge base of maximally differentiated concepts by selecting training examples in a prudent way. Note that positive examples represent information about intra-concept similarities, whereas negative examples represent information about inter-concept differentiation. Hence to achieve a set of maximally differentiated concept descriptions, one can employ all positive examples of a given concept as negative examples for all other disjoint concepts. Furthermore one can employ any given ontological concept as negative example for any other disjoint ontological concept.[4] Since the second learning objective presented leads us to select the most general concept hypothesis as the logical conjunction from the general boundary set GB, all inter-concept discriminating attributes introduced through negative examples will be represented. This provides a theoretic foundation for learning a knowledge base of maximally differentiated concept descriptions.

V. Experimental results

A. Application to the eTRIMS domain

In the context of the eTRIMs project, the concept learning approach presented here has been applied to the domain of terrestrial views of building facades. Typical aggregates of this domain are window arrays (consisting of aligned and regularly spaced windows as parts), balconies (consisting of railing, door and optional windows) or entrances (consisting of a door, stairs and ground).

To conduct the actual training sequence, positive learning examples are directly extracted from annotated pictures. An enriched instance description of the positive example is generated from the information contained in the annotation of the aggregate parts.

We automatically generate negative examples from annotated pictures by selecting random sets of parts. To be precise, we select a negative example N as any set of annotated objects that is not a subset or equal to a positive example P in the same picture. This requires, of course, that positive examples are annotated to their maximal extent. Following Winston's insight about "near-miss" examples [19], one can assume a negative example N to be most useful if it differs from a positive example P as little as possible. Hence an ideal negative example differs from a positive example only in one discriminating attribute. This kind of negative example leads to the generation of a most general concept description which is only specialised to exclude the attribute value of the discriminating attribute in the negative example. A straightforward approach to generate negative examples with near-miss properties is to define a distance measure d for N and P, to randomly generate possible training examples $N_1..N_n$ and finally choose the examples minimizing d. This approach is computationally inexpensive as the distance measure from N_i to P is available at nearly no cost (compared to the cost of the training procedure). Hence a large number of negative examples can be evaluated for their near-miss properties.

For a typical training sequence, about 10 to 15 positive examples are used. Since negative examples have stronger concept differentiation qualities, we apply about 100 to 300, keeping a ratio between positive and negative examples of 1/10 to 1/20.

B. Evaluation

As an example result of the learning process we present the General Boundary conjunction hypothesis h_g for the aggregate

[4] A formal description of specialization through concepts is omitted, but very similar to specialization through examples (II-C)

”Window Array”, learnt from 13 annotated positive examples and 260 generated negative examples:

Size and configuration

```
Aggregate Width =                          ]549..INF] cm
Aggregate Height =                         [0..200[ cm
Parts Width =                              [0..INF] cm
Parts Height =                             [0..INF] cm
Parts Top-Left-X Variability =             ]131..INF] cm
Parts Top-Left-Y Variability =             [0..33[ cm
Parts Bottom-Right-X Variability =         ]115..INF] cm
Parts Bottom-Right-Y Variability =         [0..9[ cm
```

Composition

```
Has-Parts =                                [3..INF]
       window =                            [3..INF]
       door =                              [0..0]
Part-Of =                                  [1..1]
       facade =                            [0..1]
       roof =                              [0..1]
```

Symbolic attributes

```
Shape =                                    { Elongated-X }
```

Attribute predicates

```
Fuzzy-Equal                                (top-left-y)
Fuzzy-Equal                                (bottom-right-y)
Fuzzy-Equal                                (parts-height)
Fuzzy-Equal                                (parts-dist-x)
Value-Equal                                (parts-type)
```

Internal spatial relations

```
(window000) LeftNeighbourOf [132..324] cm (window001)
(window000) LeftOf [339..649] cm (window002)
(window001) LeftNeighbourOf [206..325] cm (window002)
(window001) RightNeighbourOf [132..324] cm (window000)
(window002) RightNeighbourOf [206..325] cm (window001)
(window002) RightOf [339..649] cm (window000)
```

External spatial relations

```
(concept013) BelowOf [44..1865] cm (sky020)
(sky020) AboveOf [44..1865] cm (concept013)
```

TABLE II

LEARNT AGGREGATE DESCRIPTION ”WINDOW ARRAY”

A preliminary kind of evaluation is to test learnt concepts on instances from annotated images which have not been used for training. Table III shows false negative recognitions on these instances. In Table IV we evaluated the number of false positive recognitions and added additional random sets of parts to the test set. A large scale evaluation and comparison with known classification methods will be carried out soon (albeit the fact that most classfication approaches are only able to learn a single classifier representation, e.g. [20], [21]).

To give an intuition of the learning result for a whole knowledge base, we present a graphical representation of the knowledge base learnt for the eTRIMS domain (Figure 2).

A learnt knowledge base can also be evaluated actively using the interpretation facilities of the SCENIC system. Its scene interpretation process is based on the hypothesise-and-test

Aggregate	Instances	h_s Detect. / Succ.	h_g Detect. / Succ.
”window array”	18	16 / 0.88	17 / 0.94
”balcony”	14	11 / 0.79	14 / 1.00
”entrance”	9	8 / 0.88	9 / 1.00

TABLE III

FALSE NEGATIVE RECOGNITIONS FOR h_s AND h_g

Aggregate	Instances	h_s Detect. / Succ.	h_g Detect. / Succ.
”window array”	54	0 / 1.00	0 / 1.00
”balcony”	61	0 / 1.00	0 / 1.00
”entrance”	68	1 / 0.99	3 / 0.96

TABLE IV

FALSE POSITIVE RECOGNITIONS FOR h_s AND h_g

paradigm. Hypotheses are posed mainly through part-whole-reasoning, which emphasises the role of conceptual aggregate descriptions. Fig. 3 shows the result of an interpretation process applying the learnt concept description in Table II to an image, where scene objects have been automatically detected ([22], [23]). The interpretation system tries to interpret the scene by finding object aggregations based on the detected scene objects and the ontological aggregate descriptions in the knowledge base. SCENIC infers four instances of the window array concept and poses four additional window hypotheses.

C. Feedback learning

Automatic refinement of learnt concepts is possible through an interpretation process applying them to annotated pictures which have not been used for training. The results of the interpretation process are automatically evaluated against the ground truth given by the annotation.

There are two possible cases of misinterpretation:

- An annotated instance of the concept to be evaluated is not recognised.
- A set of annotated objects is wrongly interpreted as an instance of the aggregate concept to be evaluated.

In the case of false negative recognition, the learnt concept description is too specific. As a feedback step, the unrecognised aggregate instance is introduced to the learning process as a

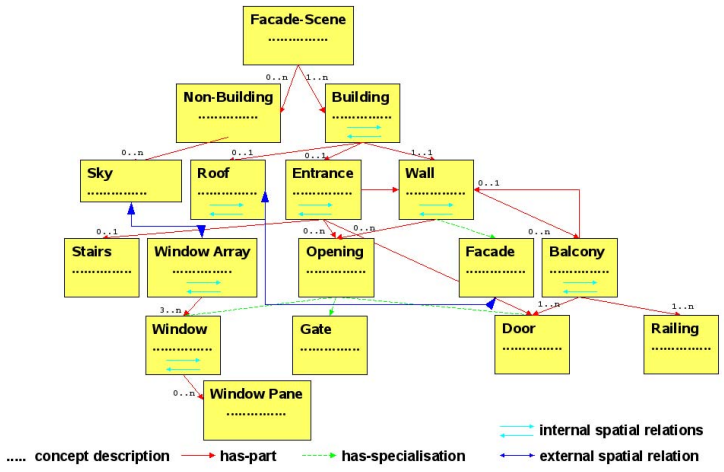

Fig. 2. Learnt building domain ontology

positive example, generalising the learnt concept description. In the case of false positive recognition, the learnt concept description is too general. Therefore the set of misinterpreted objects is introduced to the learning module as a negative example. For both cases another misclassification of the particular instance becomes impossible, regardless which concept hypothesis is chosen from VS after feedback learning.

VI. Conclusions

We have shown that conceptual descriptions for real-world knowledge-based scene interpretation can be obtained by Version Space Learning. A concept description language has been presented which allows to express quantitative as well as qualitative attributes and relations suitable for the description of ontological concepts. Employing Version Space Learning and our concept representation we are able to constitute a comprehensive knowledge base of ontological concepts with a multitude of relations between them. Additionally, the learning process gives us means to assign confidence values to learnt concepts. The resulting knowledge base is well adapted for later interpretation. Novel results also pertain to concept selection and concept differentiation for a conceptual knowledge base. Version Space Learning can be used to obtain maximally differentiated concepts. The success of learning has been demonstrated by simple evaluation and scene interpretation experiments employing the learnt concepts. By making use of annotated images, an automatic feedback learning cycle can be entered where wrong interpretations serve as correcting examples. An extended evaluation using a database of several hundred annotated facade images will be carried out soon.

Version Space Learning is attractive for stepwise extension and refinement of conceptual knowledge bases because individual examples count and mistakes can be easily corrected by feedback learning. As a drawback, standard Version Space Learning is highly sensitive to bad teaching. A single inconsistent example, wrongly annotated as positive or negative, may cause the version space to collapse. To cope with this problem we are developing an unsupervised preprocessing step, clustering the training examples and detecting outliers. In fact, this clustering approach gives us the ability to introduce examples without a-priori classification. This yields an approach to semi-supervised learning of ontological concept descriptions.

Another interesting topic for further research is to use the knowledge about near-miss properties of negative examples to derive an approach to Active Learning. If a model of best near-miss examples can be generated from a given concept description, the learner will be able to actively choose appropriate negative examples to learn a most discriminating concept description. In fact, the learner will transfer his knowledge about intra-concept similarity to derive a model for inter-concept discrimination.

Fig. 3. Detected scene objects and SCENIC interpretation result

References

[1] T.M. Mitchell, "Version spaces: A candidate elimination approach for rule learning", Proc. of the International Joint Conference on Artificial Intelligence, pp. 305–310, 1977.

[2] T.M. Mitchell, "Version Spaces: An Approach to Concept Learning", PhD thesis, Stanford University, Cambridge, MA, 1978.

[3] T.M. Mitchell, "The need for biases in learning generalizations", Technical Report CBM-TR-117, New Brunswick, New Jersey, 1980.

[4] B, Neumann, "A Conceptual Framework for High-level Vision", Technical report FBI-HH-B-241/02, Universität Hamburg, 2002.

[5] K. Sage, J. Howell, H. Buxton, "Recognition of Action", Activity and Behaviour in the ActIPret Project, Künstliche Intelligenz, 2/2005, BöttcherIT Verlag, Bremen, pp. 30–33.

[6] K. Murphy, A. Torralba, and W. T. Freeman, "Using the forest to see the trees: A graphical model relating features, objects, and scenes", Proc. of Neural Information Processing Systems, 2003.

[7] M. Boutell, J. Luo, "Scene parsing using region-based generative models", IEEE Transactions on Multimedia 9(1), pp. 136-146, 2007.

[8] L. Hotz, B. Neumann, "Scene Interpretation as a Configuration Task", Künstliche Intelligenz, 3/2005, BöttcherIT Verlag, Bremen, pp. 59–65.

[9] H. Hirsh, "Polynomial-Time Learning with Version Spaces", National Conference on Artificial Intelligence, pp. 117–122, 1992.

[10] H. Hirsh, N. Mishra, L. Pitt, "Version Spaces without boundary sets", Proc. AAAI-97, pp. 491-496, 1997.

[11] M. Sebag, "Using Constraints to Building Version Spaces", Proc. of the 7th European Conference on Machine Learning, pp. 257-271, 1994.

[12] T.-P. Hong, S.-S. Tseng, "A Generalized Version Space Learning Algorithm for Noisy and Uncertain Data", IEEE Transactions on Knowledge And Data Engineering, Vol. 9, No. 2, 1997.

[13] L. De Raedt, S. Kramer, "The Levelwise Version Space Algorithm and its Application to Molecular Fragment Finding", Proc. of the International Joint Conference on Artificial Intelligence, pp. 853–862, 2001.

[14] R.S. Michalski, "A Theory and Methodology of Inductive Learning", Machine Learning - An Artificial Intelligence Approach, pp. 83–143, 1983.

[15] B. Ganter, R. Wille, "Formal Concept Analysis - Mathematical Foundations", Springer Verlag, 1999.

[16] D. Haussler, "Quantifying inductive bias: AI learning algorithms and Valiant's learning framework", Artificial Intelligence 36, pp. 177–221, 1988.

[17] S.W. Norton, H. Hirsh, "Classifier learning from noisy data as reasoning under uncertainty", Proc. of the National Conference on Artificial Intelligence, 1992.

[18] S.W. Norton, H. Hirsh, "Learning DNF via probabilistic evidence combination", Machine Learning: Proc. of the Seventh Internation Conference, 1990.

[19] P.H. Winston, "Learning structural descriptions from examples", The psychology of computer vision, pp. 157–209, 1975.

[20] P. Mulhem, W.K. Leow, Y.K. Lee, "Fuzzy conceptual graphs for matching images of natural scenes", Proceedings of International Joint Conference on Artificial Intelligence, pages 1397-1404, 2001.

[21] J-H. Lim, Q. Tian, P. Mulhem, "Home Photo Content Modeling for Personalized Event-Based Retrieval", IEEE Multimedia, vol. 10, no. 4, pp. 28–37, October/December, 2003.

[22] J. Šochman, J. Matas, "WaldBoost - Learning for Time Constrained Sequential Detection", Proc. of the Conference on Computer Vision and Pattern Recognition, pp. 150–157, 2005.

[23] L. Hotz, B. Neumann, K. Terzic, J. Šochman, "Feedback between Low-Level and High-Level Image Processing", eTRIMS Project Deliverable D2.4, 2007.