Ontology Based Semantic Image Interpretation

Ivan Donadello[1,2]*

[1] Fondazione Bruno Kessler, Via Sommarive 18, I-38123, Trento, Italy
[2] DISI, University of Trento, Via Sommarive 9, I-38123, Trento, Italy
donadello@fbk.eu

Abstract. Semantic image interpretation (SII) leverages Semantic Web ontologies for generating a mathematical structure that describes the content of images. SII algorithms consider the ontologies only in a late phase of the SII process to enrich these structures. In this research proposal we study a well-founded framework that combines logical knowledge with low-level image features in the early phase of SII. The image content is represented with a partial model of an ontology. Each element of the partial model is grounded to a set of segments of the image. Moreover, we propose an approximate algorithm that searches for the most plausible partial model. The comparison of our method with a knowledge-blind baseline shows that the use of ontologies significantly improves the results.

1 Introduction

Semantic image interpretation (SII) is the task of generating a semantically rich structure that describes the content of an image [8]. This structure is both human and machine understandable and can be encoded by using the Semantic Web (SW) language RDF. The first advantage is that RDF enables the enrichment of the semantic content of images with SW resources, the second one is that an RDF based description of images enables content-based image retrieval via query languages like SPARQL.

The main challenge in SII is bridging the so called *semantic gap* [3], which is the complex correlation between low-level image features and high-level semantic concepts. High-level knowledge plays a key role in bridging the semantic gap [17,18]. This knowledge can be found in the ontologies provided by the SW.

Most of the current approaches to SII exploit ontologies at a later stage when some hypothesis (a geometric description of the objects and their spatial relations) of the image content have already been formulated by a bottom-up approach (see for instance [13,15,17,11,12,3,6,1]). In these cases background knowledge is exploited to check the consistency of the output and/or to infer new facts. These works do not consider uncertainty coming from the low-level image analysis or require a set of DL rules for defining what is abducible, which need to be manually crafted.

In this research proposal we study a general framework for SII that allows the integration of ontologies with low-level image features. The framework takes as input the ontology and exploits it in the process of image interpretation. The output is a description of the content of an image in terms of a (most plausible) partial logical model

* I thank my advisor Luciano Serafini for his precious help, suggestions and patience.

of the ontology [15]. Instead of lifting up low-level features into a logical form using concrete domain (as in [11]) we proceed in the opposite direction, by compiling down the background knowledge into low-level features. This allows us a more flexible inference in processing numeric information and to use simpler, and more efficient, logical reasoners for the semantic part. This partial model is generated by using optimisation methods (e.g. clustering) that integrate numeric and logical information. Our contribution is a formal framework for SII that integrates low-level features and logical axioms. Moreover, we developed an early prototype and we evaluated it, with promising results, on the task of detecting complex objects starting from the presence of their parts [5].

2 Theoretical framework

The proposed framework takes as input a *labelled picture* that is a picture partitioned into segments (regions of pixels) using a semantic segmentation algorithm [4,7]. Each segment has a set of weighted labels that represent the level of confidence of the semantic segmentation. Labels are taken from the signature Σ which is the alphabet of the ontology. A labelled picture is a pair $\mathcal{P} = \langle S, L \rangle$ where $S = \{s_1 \ldots, s_n\}$ is a set of segments of the picture $\mathcal{P}$, and L is a function that associates to each segment $s \in S$ a set $L(s)$ of weighted labels $\langle l, w \rangle \in \Sigma \times (0, 1]$.

In this research proposal we study a method for discovering new objects (e.g., composite objects) and relations between objects by exploiting low-level image features and a Description Logic (DL) [2] ontology. The ontology has the classical signature $\Sigma = \Sigma_C \uplus \Sigma_R \uplus \Sigma_I$ of symbols for concepts, relations and individuals respectively. We adopt the standard definitions for syntax and semantics of DL[3] . An ontology $\mathcal{O}$ on Σ is a set of DL axioms. An interpretation of a DL signature Σ is a pair $\mathcal{I} = \left\langle \Delta^{\mathcal{I}}, \cdot^{\mathcal{I}} \right\rangle$, where $\Delta^{\mathcal{I}}$ is a non empty set and $\cdot^{\mathcal{I}}$ is a function that interprets the symbols of Σ in $\Delta^{\mathcal{I}}$. $\mathcal{I}$ is a *model* of an ontology $\mathcal{O}$ if it satisfies all the axioms in $\mathcal{O}$. The axioms of the ontology are constraints on the states of the world. A picture, however, provides only a partial view of the state of the world, indeed, it could show a person with only one (visible) leg. Therefore, the content of a picture is not isomorphic to a model, as a model could contain objects not appearing in the picture (the invisible leg). The content of a picture should instead be represented as a *partial model*[4].

Definition 1 (Partial model). *Let $\mathcal{I}$ and $\mathcal{I}'$ be two interpretations of the signatures Σ and Σ' respectively, with $\Sigma \subseteq \Sigma'$; $\mathcal{I}'$ is an* extension of *$\mathcal{I}$, or equivalently $\mathcal{I}'$* extends *$\mathcal{I}$, if $\Delta^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}'}$, $a^{\mathcal{I}} = a^{\mathcal{I}'}$, $C^{\mathcal{I}} = C^{\mathcal{I}'} \cap \Delta^{\mathcal{I}}$ and $R^{\mathcal{I}} = R^{\mathcal{I}'} \cap \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$, for all $a \in \Sigma_I$, $C \in \Sigma_C$ and $R \in \Sigma_R$. $\mathcal{I}_p$ is a* partial model *for a ontology $\mathcal{O}$, in symbols $\mathcal{I}_p \models_p \mathcal{O}$, if there is a model $\mathcal{I}$ of $\mathcal{O}$ that extends $\mathcal{I}_p$.*

In this framework the use of DL ontologies is twofold: first they are a terminological source for labelled pictures, second the DL inference services are exploited to check if an interpretation is a partial model and thus inferring new facts. The semantic interpretation of a picture is a partial model plus an alignment, called *grounding*, of every element of $\Delta^{\mathcal{I}_p}$ with the segments of the picture.

[3] In this paper we use the $\mathcal{SHIQ}$ DL.

[4] This intuition was introduced in [15], our formalization however is slightly different.

Definition 2 (Semantically interpreted picture). *Given an ontology $\mathcal{O}$ with signature Σ and a labelled picture $\mathcal{P} = \langle S, L \rangle$, a* semantically interpreted picture *is a triple $\mathbb{S} = (\mathcal{P}, \mathcal{I}_p, \mathcal{G})_\mathcal{O}$ where:*

- *$\mathcal{I}_p = \langle \Delta^{\mathcal{I}_p}, \cdot^{\mathcal{I}_p} \rangle$ is a partial model of $\mathcal{O}$;*
- *$\mathcal{G} \subseteq \Delta^{\mathcal{I}_p} \times S$ is a left-total and surjective relation called* grounding relation*: if $\langle d, s \rangle \in \mathcal{G}$ then there exists an $l \in L(s)$ such that:*
  1. *if $l \in \Sigma_C$ then $d \in l^{\mathcal{I}_p}$;*
  2. *if $l \in \Sigma_I$ then $d = l^{\mathcal{I}_p}$;*
  3. *if $l \in \Sigma_R$ then $\langle d, d' \rangle \in R^\mathcal{I}$ or $\langle d', d \rangle \in R^\mathcal{I}$ for some $d' \in \Delta^\mathcal{I}$.*

The grounding *of every $d \in \Delta^{\mathcal{I}_p}$, denoted by $\mathcal{G}(d)$, is the set $\{s \in S \mid \langle d, s \rangle \in \mathcal{G}\}$.*

There are many possible explanations of the picture content, thus there are many partial models describing a picture via a grounding relation. We define a cost function $\mathcal{S}$ that assigns a cost to a partial model based on its adherence to the image content: the higher the adherence the lower the cost. The *most plausible partial model* $\mathcal{I}_p^*$ is the partial model that minimizes $\mathcal{S}$, in symbols:

$$\mathcal{I}_p^* = \underset{\substack{\mathcal{I}_p \models_p \mathcal{O} \\ \mathcal{G} \subseteq \Delta^{\mathcal{I}_p} \times S}}{\operatorname{argmin}} \ \mathcal{S}(\mathcal{P}, \mathcal{I}_p, \mathcal{G})_\mathcal{O} \tag{1}$$

The definition of $\mathcal{S}$ has to take into account low-level features of the segments and high-level semantic features of the partial model derivable form the ontology. Intuitively the cost function measures the semantic gap between the two types of features.

Definition 3 (Semantic image interpretation problem). *Given an ontology $\mathcal{O}$ and a labelled picture $\mathcal{P}$, a cost function $\mathcal{S}$, the* semantic image interpretation problem *is the construction of a semantically interpreted picture $\mathbb{S} = (\mathcal{P}, \mathcal{I}_p, \mathcal{G})_\mathcal{O}$ that minimizes $\mathcal{S}$.*

3 Method

In this proposal we restrict to the recognition of complex objects from their parts. For example, given a labelled picture where only some parts of a man (the legs, one arm and the head) and of a horse (the legs, the muzzle and the tail) are labelled we want to infer the presence of some logical individuals with their classes (man and horse respectively). These individuals are linked with their parts through the partOf relation. This can be seen as a clustering problem and we specify the cost function in terms of clustering optimisation. The parts (simple objects) are the input of the clustering problem whereas a single cluster contains the parts of a composite object. In addition, the parts to cluster are the individuals $d \in \Delta^{\mathcal{I}_p}$ with the following features:

- a set of low-level image features extracted from $\mathcal{G}(d)$, the grounding of d;
- a set of semantic features corresponding the most specific concepts extracted from the set $\{C \in \Sigma_C \mid d \in C^{\mathcal{I}_p}\}$ assigned to d by $\mathcal{I}_p$.

We use the centroid of $\mathcal{G}(d)$ as a numeric feature but the approach can be generalised to other features. Clustering algorithms are based on some distance between the input elements defined in terms of their features. Let $\delta_{\mathcal{G}}(d, d')$ be the Euclidean distance of the centroids of $\mathcal{G}(d)$ and $\mathcal{G}(d')$, $\delta^s_{\mathcal{O}}(d, d')$ a semantic distance between simple objects and $\delta^c_{\mathcal{O}}(d, d')$ a semantic distance between a simple object and its corresponding composite object. We define the cost function as the quality measure of the clustering:

$$\mathcal{S}(\langle \mathcal{P}, \mathcal{I}_p, \mathcal{G} \rangle_{\mathcal{O}}) = \alpha \left(\sum_{d,d' \in (\exists \mathsf{hasPart}.\top)^{\mathcal{I}_p}} \delta_{\mathcal{G}}(d, d') \right) +$$

$$(1-\alpha) \left(\sum_{\substack{\langle d', d \rangle \in \mathsf{partOf}^{\mathcal{I}_p} \\ \langle d'', d \rangle \in \mathsf{partOf}^{\mathcal{I}_p}}} (\delta_{\mathcal{G}}(d', d'') + \delta^s_{\mathcal{O}}(d', d'')) + \sum_{\langle d', d \rangle \in \mathsf{partOf}^{\mathcal{I}_p}} (\delta_{\mathcal{G}}(d', d) + \delta^c_{\mathcal{O}}(d', d)) \right).$$

Following [9], the first component of the above equation measures the centroid distance between the composite objects (inter-cluster distance). The second component estimates the distance between the elements of each single cluster (intra-cluster distance).

Minimising analytically the above equation is rather complex, thus we developed an iterative algorithm that at each loop groups the several parts of a composite object approximating the cost function. If the grouping is not a partial model the algorithm enters in the next loop and selects another clustering. In the first step our algorithm generates an initial partial model $\mathcal{I}_p$ from $\mathcal{P} = \langle S, L \rangle$ where $\Delta^{\mathcal{I}_p}$ contains an element d_s for every segment $s \in S$ and any concept C in the labelled picture is interpreted as $C^{\mathcal{I}_p} = \{d_s | C \in L(s)\}$. The grounding $\mathcal{G}$ is the set of pair $\langle d_s, s \rangle$. Then, the algorithm enters in a loop where a non-parametric clustering procedure [10] clusters the input elements $d \in \Delta^{\mathcal{I}_p}$ by using their numeric and semantic features according to $\delta_{\mathcal{G}}$ and $\delta^s_{\mathcal{O}}$. Each cluster cl corresponds to a composite object d_{cl} which is introduced in $\mathcal{I}_p$ and is connected via the $\mathsf{hasPart}$ relation to the elements of cl. We predict the type of this new individual via abductive reasoning: the type is the ontology concept that shares the maximum number of parts with the elements of the cluster. For example, if we cluster some elements of type Tail, Muzzle and Arm an abducted ontology concept will be Horse. These new facts are introduced in $\mathcal{I}_p$ and the algorithm checks if $\mathcal{I}_p$ is a partial model of $\mathcal{O}$ by using a DL reasoner (Pellet [16]). If true the algorithm returns $\mathcal{I}_p$ otherwise it extends the input elements with a set of consistency features that encode information about the inconsistency of $\mathcal{I}_p$. These features tend to separate (resp. join) the segments that have been joined (resp. separated) in the previous clustering. The cluster of our example is inconsistent because a horse does not have arms. Then the algorithm returns at the beginning of the loop.

4 Evaluation

To evaluate our approach we created, by using LABELME [14], a dataset of 204 labelled pictures. For each picture we manually annotated simple objects, composite objects and

Table 1. Performance of the proposed algorithm for SII and comparison with the baseline. The reported data are the average of the three measures on each single picture.

	$prec_{\mathrm{GRP}}$	rec_{GRP}	$F1_{\mathrm{GRP}}$	$prec_{\mathrm{COP}}$	rec_{COP}	$F1_{\mathrm{COP}}$
SII	**0.61**	**0.89**	**0.67**	**0.73**	**0.75**	**0.74**
Baseline	0.45	0.71	0.48	0.66	0.69	0.66

their part-whole relations[5]. We also created a simple ontology[6] with a basic formalisation of meronymy in the domains of: houses, trees, people, and street vehicles. We built a ground truth by associating every single labelled picture $\mathcal{P}$ to its partial model encoded in an ABox $A_{\mathcal{P}}$. The partial model returned by our algorithm is encoded in the $A^*_{\mathcal{P}}$ ABox, in order to compare $A_{\mathcal{P}}$ with $A^*_{\mathcal{P}}$ we define the following two measures.

Grouping (GRP): this measure expresses how good is our algorithm at grouping parts of the same composite object. We define precision, recall and F1 measure on the set of siblings (the parts of the same composite object): $\mathsf{sibl}(\mathcal{A}) = \{\langle d, d'\rangle \mid \exists d'' : \mathsf{partOf}(d, d''), \mathsf{partOf}(d', d'') \in \mathcal{A}\}$. Thus:

$$prec_{\mathrm{GRP}}(\mathcal{P}) = \frac{|\mathsf{sibl}(\mathcal{A}_{\mathcal{P}}) \cap \mathsf{sibl}(\mathcal{A}^*_{\mathcal{P}})|}{|\mathsf{sibl}(\mathcal{A}^*_{\mathcal{P}})|} \qquad rec_{\mathrm{GRP}}(\mathcal{P}) = \frac{|\mathsf{sibl}(\mathcal{A}_{\mathcal{P}}) \cap \mathsf{sibl}(\mathcal{A}^*_{\mathcal{P}})|}{|\mathsf{sibl}(\mathcal{A}_{\mathcal{P}})|}$$

Complex-object prediction (COP): this measure expresses how good is our algorithm at predicting the type of the composite object. We define precision, recall and F1 measure on the types of the composite object each part is assigned to: $\mathsf{ptype}(\mathcal{A}) = \{\langle d, \mathsf{C}\rangle \mid \exists d' : \{\mathsf{partOf}(d, d'), \mathsf{C}(d')\} \subset \mathcal{A}\}$. Thus:

$$prec_{\mathrm{COP}}(\mathcal{P}) = \frac{|\mathsf{ptype}(\mathcal{A}_{\mathcal{P}}) \cap \mathsf{ptype}(\mathcal{A}^*_{\mathcal{P}})|}{|\mathsf{ptype}(\mathcal{A}^*_{\mathcal{P}})|} \quad rec_{\mathrm{COP}}(\mathcal{P}) = \frac{|\mathsf{ptype}(\mathcal{A}_{\mathcal{P}}) \cap \mathsf{ptype}(\mathcal{A}^*_{\mathcal{P}})|}{|\mathsf{ptype}(\mathcal{A}_{\mathcal{P}})|}$$

To measure how the semantics improves the recognition of composite objects from their parts we implemented a baseline that clusters without semantic features, see Table 1. We can see that the explicit use of semantic knowledge via semantic distance, abductive and deductive reasoning improves the baseline that relies only on numeric features.

5 Conclusions

We proposed a well-founded and general framework for SII that integrates symbolic information of an ontology with low-level numeric features of a picture. An image is interpreted as a (most plausible) partial model of an ontology that allows the query about the semantic content. We applied the framework to the specific task of recognizing composite objects from their parts. The evaluation shows good results and the injection of semantic knowledge improves the performance with respect to a semantically-blind baseline. As future work, we want to extend our evaluation by using more low-level features, by studying other relations and by using a semantic segmentation algorithm as source of labelled pictures.

[5] An example of labelled picture is available at `http://bit.ly/1DXZxic`

[6] The ontology is available at `http://bit.ly/1AruGh0`

References

1. Atif, J., Hudelot, C., Bloch, I.: Explanatory reasoning for image understanding using formal concept analysis and description logics. Systems, Man, and Cybernetics: Systems, IEEE Transactions on 44(5), 552–570 (May 2014)
2. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, New York, NY, USA (2003)
3. Bannour, H., Hudelot, C.: Towards ontologies for image interpretation and annotation. In: Martinez, J.M. (ed.) 9th International Workshop on Content-Based Multimedia Indexing, CBMI 2011, Madrid, Spain, June 13-15, 2011. pp. 211–216. IEEE (2011)
4. Carreira, J., Caseiro, R., Batista, J., Sminchisescu, C.: Semantic segmentation with second-order pooling. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012. LNCS, Springer Berlin Heidelberg (2012)
5. Donadello, I., Serafini, L.: Mixing low-level and semantic features for image interpretation. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) Computer Vision - ECCV 2014 Workshops. LNCS, Springer International Publishing (2014), best paper award.
6. Espinosa, S., Kaya, A., Möller, R.: Logical formalization of multimedia interpretation. In: Paliouras, G., Spyropoulos, C., Tsatsaronis, G. (eds.) Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, Lecture Notes in Computer Science, vol. 6050, pp. 110–133. Springer Berlin Heidelberg (2011)
7. Gould, S., Zhao, J., He, X., Zhang, Y.: Superpixel graph label transfer with learned distance metric. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. Lecture Notes in Computer Science, Springer International Publishing (2014)
8. Hudelot, C., Maillot, N., Thonnat, M.: Symbol grounding for semantic image interpretation: From image data to semantics. In: Proc. of the 10th IEEE Intl. Conf. on Computer Vision Workshops. ICCVW '05, IEEE Computer Society (2005)
9. Jung, Y., Park, H., Du, D.Z., Drake, B.L.: A decision criterion for the optimal number of clusters in hierarchical clustering. Journal of Global Optimization 25(1), 91–111 (2003)
10. Kohonen, T.: The self-organizing map. Proc. of the IEEE 78(9), 1464–1480 (Sep 1990)
11. Neumann, B., Möller, R.: On scene interpretation with description logics. Image and Vision Computing 26(1), 82 – 101 (2008), cognitive Vision-Special Issue
12. Peraldi, I.S.E., Kaya, A., Möller, R.: Formalizing multimedia interpretation based on abduction over description logic aboxes. In: Proc. of the 22nd Intl. Workshop on Description Logics (DL 2009). CEUR Workshop Proceedings, vol. 477. CEUR-WS.org (2009)
13. Reiter, R., Mackworth, A.K.: A logical framework for depiction and image interpretation. Artificial Intelligence 41(2), 125–155 (1989)
14. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: A database and web-based tool for image annotation. Int. J. Comput. Vision 77(1-3), 157–173 (May 2008)
15. Schroder, C., Neumann, B.: On the logics of image interpretation: model-construction in a formal knowledge-representation framework. In: Image Processing, 1996. Proceedings., Int. Conf. on. vol. 1, pp. 785–788 (Sep 1996)
16. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical owl-dl reasoner. Web Semant. 5(2), 51–53 (Jun 2007)
17. Town, C.: Ontological inference for image and video analysis. Mach. Vision Appl. 17(2), 94–115 (Apr 2006)
18. Yuille, A., Oliva, A.: Frontiers in computer vision: Nsf white paper (November 2010), `http://www.frontiersincomputervision.com/WhitePaperInvite.pdf`