A Model for the Semantic Annotation of Images

Master’s Thesis

Presented to
The Faculty of the Graduate School of Arts and Sciences
Brandeis University
Department of Computer Science
James Pustejovsky, Advisor

In Partial Fulfillment
of the Requirements for the Degree
Master of Arts
in
Computational Linguistics

by
Julia Bosque Gil
August, 2014
Acknowledgements

I would like to offer my special thanks to Dr. James Pustejovsky, my supervisor, for suggesting me this topic of research and guiding me when I was confused and overwhelmed by the amount of literature on events in linguistics, natural language processing, and ontology engineering. His comments and suggestions about how to endeavor in this project and about the model specification have been extremely helpful and very much appreciated. I also wish to thank Dr. Asunció Gómez Pérez, Dr. Elena Montiel Ponsoda, Dr. Guadalupe Aguado de Cea and Dr. Ricardo Mairal Usón for their valuable advice before I even started the thesis. Although I approached a different topic in the end, their insights on event detection projects were greatly appreciated and will be taken into consideration for future work.
Abstract

A Model for the Semantic Annotation of Images

A thesis presented to the Department of Computer Science

Graduate School of Arts and Sciences of
Brandeis University, Waltham, Massachusetts

by Julia Bosque Gil

The idea behind the task of semantic annotation of images rests on the ability to retrieve an image using a query by concept, and not a query by keyword match in the text surrounding the image or a query by low-level features. The term semantic gap refers to the breach between the content the user wants to find when she searches for an image and the actual low-level features corresponding to that content. Semantic annotation approaches focus on bridging this gap by recording content-relevant information, for instance, the categories to which objects in the image belong (e.g. dog, person), the event depicted in the scene (e.g. sleeping) and the location and setting of the action. This thesis proposes a model for complexly structured semantic annotation of images depicting events and activities that incorporates notions from text annotation schemes such as SpatialML (Mani et al. 2010) and ISO-Space (Pustejovsky et al. 2013). An annotation that characterizes events, the semantic roles of participants, the motion and visual characteristics of figures, as well as details concerning the location and time, would play a significant role in structured image retrieval, and a mapping of annotated semantic entities and the image’s low-level features would likely assist the event recognition and description generation tasks.
## Contents

Acknowledgements .......................... ii

Abstract .................................... iii

1 Introduction ................................ 1

2 Image Annotation ............................ 5
   2.1 Overview ................................ 5
   2.2 Keyword Annotation ....................... 6
   2.3 Annotation within MPEG-7 ................. 8
      2.3.1 Semantic Descriptors .................... 9
      2.3.2 Caliph and Emir ......................... 11
   2.4 Annotation Using Ontologies ............... 14
   2.5 Description Generation ..................... 16
   2.6 Annotation of Spatial Information in Natural Language Text ............ 18
      2.6.1 SpatialML ............................... 18
      2.6.2 ISO-Space ................................ 20
      2.6.3 Spatial Role Labeling ................... 22

3 A Model for the Semantic Annotation of Images .......................... 23
   3.1 Overview ................................ 23
   3.2 Model ................................... 27
      3.2.1 Conceptual Schema ...................... 27
      3.2.2 Elements ............................... 28
         3.2.2.1 IMAGE ............................... 28
         3.2.2.2 FIGURE .............................. 29
         3.2.2.3 OFigure ............................ 32
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.2.4</td>
<td>Event</td>
<td>32</td>
</tr>
<tr>
<td>3.2.2.5</td>
<td>Setting</td>
<td>34</td>
</tr>
<tr>
<td>3.2.2.6</td>
<td>Time</td>
<td>35</td>
</tr>
<tr>
<td>3.2.2.7</td>
<td>Rolelink</td>
<td>36</td>
</tr>
<tr>
<td>3.2.2.8</td>
<td>Holdinglink</td>
<td>38</td>
</tr>
<tr>
<td>3.2.2.9</td>
<td>Motionlink</td>
<td>38</td>
</tr>
<tr>
<td>3.2.2.10</td>
<td>Dirlink</td>
<td>40</td>
</tr>
<tr>
<td>3.2.2.11</td>
<td>Facelink</td>
<td>41</td>
</tr>
<tr>
<td>3.2.2.12</td>
<td>Gazelink</td>
<td>42</td>
</tr>
<tr>
<td>3.2.2.13</td>
<td>Exlink</td>
<td>42</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Annotation Examples</td>
<td>43</td>
</tr>
</tbody>
</table>

4 Conclusion | 47
## List of Figures

2.1 MPEG Description Tools .................................................. 8
2.2 MPEG-7-based Semantic Description of an Image .................... 10
2.3 Relations in MPEG-7 Semantic Annotation ............................ 11
2.4 MPEG-7 Semantic Annotation in XML .................................. 12
2.5 Sample Images from the UIUC Pascal Sentence Dataset ............... 17
2.6 Visual Dependency Representation (VDR) of an image ................. 17
2.7 SpatialML Annotation Example .......................................... 19
2.8 Spatial Role Labeling Example ......................................... 22
3.1 Conceptual Schema of the Model ....................................... 28
3.2 Annotation Example 1 ...................................................... 44
3.3 Annotation Example 2 ...................................................... 45
3.4 Annotation Example 3 ...................................................... 46
Chapter 1

Introduction

The development and emergence of new technologies over the last decades has enabled the easy generation of multimedia content and the association of information to digital images to allow for their efficient query and retrieval has grown essential. Text-based image retrieval involves a keyword matching process: given the user’s query, search engines check the image’s title, its captions, and its surrounding text for potential matches, and, if any are found, the image is retrieved as a relevant picture. This approach assumes, first, that images have an annotation of some kind or at least a text in which to perform the query, and, second, that the text of the web page on which it appears is related to the content of the image. However, these two assumptions do not always hold. During the 1990’s and due to the increase in image data in the Web and the time-consuming nature of manual annotation, content-based image retrieval approaches were proposed: images would be indexed by their visual features (dominant color, texture, shape of objects, etc.), hence ‘low-level features’, extracted through computer vision techniques (Sikora 2001). Advances in this area led to the design of standards such as the MPEG-7 format (Martinez 2004, section 2.3) to gather and store this information in a structured way to allow for its easy access and retrieval.

The idea behind the task of semantic annotation of images rests on the ability to retrieve an image using a query by concept, and not a query by keyword match in the text surrounding
the image or a query by low-level features. The term *semantic gap* refers to the breach between the content the user wants to find when she searches for an image, and the actual low-level features corresponding to that content. Semantic annotation approaches focus on bridging this gap by recording content-relevant information, for instance, the categories to which objects in the image belong (e.g. *tree, car, dog, person*), the event or situation depicted in the scene (e.g. *sleeping*), the location and the setting of the action. Efforts in automatic and manual keyword annotation or tagging, linking annotations of image content to Semantic Web ontologies, annotation using MPEG-7 semantic descriptors, or the representation of image content for the training of automatic caption generation models are current lines of work in this area. Though these approaches do capture the content of an image in different levels of detail, there is still much information not accounted for beyond the *who, what, when, where* differentiation. Textual annotation schemes that capture spatial information encoded in text, such as SpatialML (Mani et al. 2010), ISO-Space (Pustejovsky et al. 2013), and Spatial Role Labeling (Kordjamshidi et al. 2010), however, can provide new insights to this task, and are therefore included here as part of the second chapter.

This MA thesis is an attempt at semantic annotation of images depicting events and activities. The suggested model covers information about (1) the event: the type of event, any subevents involved in it, any motion triggered by it and any other event the image might denote, if it is ambiguous; (2) the participants of the event: their type of entity, their thematic roles, their appearance and their representation from the viewer’s point of view, and (3) about the setting and the time of the situation. Google’s text-based image retrieval shows that there is still much work to be done in this respect: given a query of an activity, e.g. *playing guitar*, Google returns numerous appropriate results, but more complex queries involving details about the activity and the participants of the event yield mixed or even poor results. As an example, *watering the plants* returns a fair amount of relevant pictures, whereas *watering the plants at night* only three correct photos; *throwing a ball backwards* yields only six correct images, and *singing sitting down* (*while seated, while sitting down,*
singing sitting were also tested) gives mixed results including birds perched on branches and people who are not singing. If the query includes activity keywords that also match a noun (e.g. skipping the rope – skipping rope, ironing a blanket – ironing blanket) the resulting set is a combination of images denoting activities and images denoting a single object.

An annotation schema that captured these nuances in a structured way would presumably contribute to a better image retrieval, allowing for queries by the event and semantic roles of participants, by the visual characteristics of objects, or by aspects related to the location and time. Annotated content could also be mapped to the image’s low-level feature values (e.g. if the setting is the ocean, the dominant color would be blue, and this would be already captured by a low-level feature descriptor) in order to learn visual models in the tasks of event recognition and content description generation.

The model assumes an implementation in XML, and the images presented as examples in the last section of the thesis were annotated with the text-annotation tool MAE (Stubbs 2011), slightly changed to display images on the main panel. The design of an application for image annotation that allows the user to load a DTD and an image and, additionally, annotate object regions with the appropriate model tags, in a similar way to LabelMe (Russell et al. 2008) was out of the scope of this thesis, but will be addressed in section 4. Regarding its limitations, the model does not cover the spatial configuration of objects with respect to other objects: this was already addressed in the annotation of image captions with ISO-Space (Pustejovsky and Yocum 2014), and, less deeply, in the work of Farhadi et al. (2010) and Elliott and Keller (2011, 2013) with Visual Dependency Graphs (section 2.5).

The structure of the thesis is as follows: chapter 2 provides an overview of the main approaches in semantic annotation, including tagging, MPEG-7 based annotation and annotation using ontologies, along with their advantages and limitations. Automatic caption generation is briefly introduced but not analyzed in subsection 2.5, since its ultimate goal involves automatic natural language generation and not content annotation or structuring.

\footnote{Results from Google Image Search at the time of writing, https://images.google.com/}
for a better image retrieval, in contrast the previous three lines of work. Schemes that cap-
ture spatial information in text are explained in section 2.6, and somehow provide the bridge
between semantic annotation in the computer vision community and natural language pro-
cessing. I will draw upon several of the elements of these schemes in the model presented
in chapter 3. Some of the key aspects to bear in mind when designing a model for semantic annotation are presented in that section as well, followed by the possible use cases of the model and its conceptual schema. The specification of the model constitutes the main part of chapter 3, and some annotation examples are provided afterwards. Section 4 concludes and describes directions for future work.
Chapter 2

Image Annotation

2.1 Overview

There are several approaches to the semantic annotation task. On the one hand, we have keyword annotation or tagging (Von Ahn and Dabbish 2004; Klavans et al. 2008; Von Ahn et al. 2006; Ho et al. 2009), its collaborative variant using social networks (Aurnhammer et al. 2006, Shevade and Sundaram 2005), and the advances in automatic tagging (Luo et al. 2009; Li and Fei-Fei 2007; Feng and Lapata 2010). The annotation of images with natural language descriptions is also a productive area, where schemes that capture spatial information encoded in text could play a significant role (Kordjamshidi et al. 2010; Mani et al. 2010; Pustejovsky et al. 2013; Pustejovsky and Yocum 2014). Automatic content description (Elliott and Keller 2013; Gupta 2013; Farhadi et al. 2010) is considered a language generation problem in which, given an unannotated image, the system provides a high level natural language description of its content. This involves sentence generation, in contrast to the automatic tagging, image labeling or event (action) recognition tasks, which strongly rely on computer vision techniques. On the other hand, the development of the MPEG-7 format (Multimedia Content Description Interface) (Martinez 2004) and its semantic annotation tools has inspired annotation projects that draw upon MPEG-7 Semantic Descriptors (Lux
et al. 2003; Lux and Granitzer 2005) and design ontologies that cover them (Rahman et al. 2006). Lastly, the emergence and proliferation of Semantic Web technologies and ontologies that use them has led to numerous approaches that map concepts and properties defined in an ontology to the ‘low-level’ features (color, texture, etc.) of the image or to a minimal structured annotation of its content (Hollink et al. 2003; Bloehdorn et al. 2005). This section provides a brief overview of these approaches, their advantages and limitations in image annotation.

2.2 Keyword Annotation

Image tagging involves assigning to an image labels or tags (unstructured text) that capture the content presented in it. Tags provide high-level descriptions that usually cannot be derivable from the image’s low-level features alone, with some exceptions in which the tag is grounded, e.g. \#blackandwhite (Aurnhammer et al. 2006).

Efficient image annotation is especially important in Art History. Klavans et al. (2008) develop a toolkit for catalogers that aims to make subject annotation of art works an easier task by automatically identifying potential keywords in texts about the specific work. The keywords are then mapped to different domain thesauri (e.g. Art and Architecture Thesaurus (AAT)) and geographic names databases (The Getty Thesaurus of Geographic Names (TGN)).\footnote{The mapping to domain thesauri was already proposed in Hollink et al. (2003).} The annotation of works of art follows different guidelines (e.g. Cataloging Cultural Objects Project, Baca et al. 2006) that vary depending on the institution, but they tend to capture the basic information regarding the work: title, materials, composition, format, artist, period, etc.

Games are also a useful strategy to engage people in image tagging. In Von Ahn and Dabbish 2004’s game, players are shown unannotated images and are asked to enter a label
describing their content. If both players enter the same keyword, it is associated to the 
image and the game continues. This approach tackles the problem of subjectivity and is a 
promising way of gathering reliable annotations. Ho et al. (2009) introduce a third player 
whose role is to prevent the other two from agreeing, which prevents cheating and forces the 
players to use a more diverse set of keywords. In Von Ahn et al. (2006)’s game, users have 
to locate in the image the object denoted by a given keyword in order to help collecting metata-data for object segmentation.

Collective annotation is related to tagging in the context of social networks such as 
Flickr. Tagging is necessary to easily manage and share the content a user uploads, but 
different users might use the same tag for different images, which relates to the subjective 
aspect of annotation. Shevade and Sundaram (2005)’s system incorporates computer vision 
techniques and categorized tags (Who, What, Where, When) to recommend labels to users, 
given the user’s and the group’s previous annotations, as well as concept relations among 
those keywords (ConceptNet).² Aurnhammer et al. (2006) also combine low-level features 
with keyword annotations in Flickr to show that each kind of information complements the 
other, which allows for a more accurate image retrieval.

Many of the major advances in automatic tagging come from the object recognition 
community. Automatic tagging would contribute to better image retrieval systems, as well 
as to filtering of inappropriate content and to accessibility to visually impaired users (Li 
and Fei-Fei 2007). Luo et al. (2009) map Named Entities and verbs in image captions to 
faces and poses of figures in order to train visual appearance models for event and figure 
recognition. Li and Fei-Fei (2007)’s system automatically labels event, objects and setting by 
executing an algorithm that recognizes the event in terms of objects and setting (e.g. athlete 
+ mountain = climbing), although it ignores spatial relations. Feng and Lapata (2010), on 
the other hand, identify the most representative keywords for an image given its captions 
and the text that surrounds it.

²http://conceptnet5.media.mit.edu/, last version (Speer and Havasi 2013).
Problems

Linking keywords to their corresponding low-level features is closely related to mapping concepts of an ontology to those features (section 2.4). Tagging approaches, however, do not capture any information about the specific presentation of the content: the orientation of objects relative to the viewer, spatial relations among them or the topological relations among the regions they occupy, the role they play in the event or any subevent involved in it. There is little information about how the different tags relate to each other beyond the categorization in \textit{who}, \textit{what}, \textit{where}, \textit{when} sets provided in some approaches. It is unlikely that a user searches for a random image by its author, period, or photographic or painting technique outside the context of artistic works, and the query is more likely to be concerned with image content. A keyword may suffice for simple queries, such as \textit{piano}, but more complex queries require a structured image annotation. E.g. \textit{someone crossing arms while playing piano}.

2.3 Annotation within MPEG-7

![MPEG Description Tools](image)

Figure 2.1: MPEG Description Tools (M.M. Jose, 2004)
CHAPTER 2. IMAGE ANNOTATION

Multimedia content can be viewed as a coin: on one side resides the information related to visual features, its spatial structure (or temporal, if it is a video), its decomposition in regions or shots, etc. This is the multimedia structure or the structural aspect of the multimedia file (Figure 2.1). The other side of the coin, the semantic aspect, is related to the semantics of the content of the image, and that is the side which semantic annotation addresses (Martinez 2004). The gap between both sides is called the semantic gap and the approaches in semantic annotation attempt to bridge it by capturing how low-level features of the image are related to their high-level descriptions. The structural and semantic aspect of multimedia constitute the actual content of the image (content description), perceivable by the viewer. Information about the title, the format, encoding or user rights is stored in the content management component of the multimedia (Martinez 2004).

The format MPEG-7, defined by the Moving Pictures Experts Group, specifies a set of descriptors, or representations of visual features (color layout, shape, texture, etc.) that encode information about the content description side of the multimedia file and allow for its efficient retrieval.\(^3\) MPEG-7 defines the syntax and each aspect of the features, as well as the tools to create new ones. The structure and relations between the components of a descriptor are specified by Descriptor Schemes (DS), and new Descriptors and Descriptor Schemes are defined using the Descriptor Definition Language (DDL) (Rahman et al. 2006).

2.3.1 Semantic Descriptors

One of MPEG-7’s descriptors is the semantic descriptor. It is used to describe the content of multimedia files in terms of its semantic entities such as objects, places, events, etc., the relations among them, and their attributes. All entities are part of a narrative world (Benitez et al. 2002), which is described with the Semantic Descriptor Scheme (DS). Semantic entities

---

\(^3\) The visual features represented by MPEG-7 are broadly the following: color (dominant, layout, color spaces, scalable color, group-of-frames); shape (region, contour-based, spectrum, region based, 2D-3D); texture ((non)homogenous) and motion (local, global) (Sikora 2001). For a complete list of features, see Hunter (2005).
(Figure 2.2) share some features such as labels, textual descriptions, properties, links to the media, etc., and they are captured with specialized Semantic Base DSs (the abstract class they extend): Object, AgentObject, Event, SemanticPlace, SemanticTime, Semantic State, and Concept DS (Benitez et al. 2002).  

Objects and Event DS are perceivable by the user and correspond to the event and its participants. The SemanticPlace and SemanticTime DS describe the location and time of the narrative world respectively, and can also be used to describe lengths and durations with a point of reference using the data types extent (4 miles, 3 days) and position (4 miles from NYC, the 3rd of April). The SemanticState DS describes the attributes of an entity at a given point and location in the narrative world and can reflect changes of these values in time in video data. Concepts are collections of properties and refer to any affective property of the media such as suspense or happiness (Benitez et al. 2002).

Figure 2.2: Semantic description of an image (Benitez, 2002)

Semantic relations in MPEG-7 represent how semantic entities relate to each other in the narrative world and can be divided into three groups (Figure 2.3). Narrative relations go

---

4Properties express adjectival qualities of figures, such as tall, but the MediaOccurrence element can be used instead of properties to describe the appearance of a semantic entity by associating its spatial location to visual features (Benitez et al. 2002).
back to linguistic theory and are roughly characterized in terms of thematic roles; definitive relations describe how entities are defined in terms of each other (e.g. composition, is-a relations, etc.), and spatial and temporal relations capture where semantic entities are in the segments of the image or in other semantic entities (Rahman et al. (2006)).

<table>
<thead>
<tr>
<th>Type</th>
<th>Normative relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Narrative</td>
<td>agent, agentOf, patient, patientOf, experiencer, experiencerOf, stimulus, stimulusOf,</td>
</tr>
<tr>
<td></td>
<td>causer, causerOf, goal, goalOf, beneficiary, beneficiaryOf, them, themOf, result,</td>
</tr>
<tr>
<td></td>
<td>resultOf, instrument, instrumentOf, accompanier, accompanierOf, summarizes,</td>
</tr>
<tr>
<td></td>
<td>summarizedBy, state, stateOf</td>
</tr>
<tr>
<td>Definitive</td>
<td>combination, specializes, generalizes, similar, opposite, exemplifies, exemplifiedBy,</td>
</tr>
<tr>
<td></td>
<td>interchangeable, identifier, part, partOf, contrasts, property, propertyOf, user,</td>
</tr>
<tr>
<td></td>
<td>userOf, component, componentOf, substance, substanceOf, entailment, entailmentOf,</td>
</tr>
<tr>
<td></td>
<td>manner, mannerOf, influences, dependsOn, membershipFunction</td>
</tr>
<tr>
<td>Spatial/Temporal</td>
<td>key, keyFor, annotates, annotatedBy, shows, appearsIn, reference, referenceOf,</td>
</tr>
<tr>
<td></td>
<td>quality, qualityOf, location, locationOf, source, sourceOf, destination, destinationOf,</td>
</tr>
<tr>
<td></td>
<td>path, pathOf, time, timeOf, depicts, depictedBy, represents, representedBy, context,</td>
</tr>
<tr>
<td></td>
<td>contextFor, interprets, interpretedBy</td>
</tr>
</tbody>
</table>

Figure 2.3: Relations in MPEG-7 Semantic Annotation (Rahman et al. 2006)

Structured free-text annotation is also supported with the element TextAnnotation. Text annotations can be just free-text, a structured annotation composed by Who, WhatAction, Where, When, Why and How fields, or a dependency structure annotation consisting of a dependency graph of the caption (Kosch 2013).

2.3.2 Caliph and Emir

Lux et al. (2003); Lux and Granitzer (2005) and Lux (2009)'s prototypes Caliph and Emir provide an intuitive tool for semantic image annotation and retrieval, as well as low-level feature extraction (scalable color, color layout) in terms of MPEG-7 based descriptors. Caliph
Figure 2.4: MPEG-7 Semantic Annotation (Rahman et al. 2006)
supports MPEG-7 StructuredText Annotation, but its contribution lies mainly on the semantic descriptor panel: the user can define semantic entities and use them as nodes of a graph, with the edges being the standard’s semantic relations. Retrieval with Emir takes a query that uses structured annotation, low-level features, or semantic annotation through a user-defined graph whose wildcard nodes (‘*’) function as place holders for semantic entities.

Problems

Semantic annotation with MPEG-7 semantic descriptors does capture basic information related to the event presented in the image, its participants, the setting and the time. Semantic roles are described by narrative relations, and is-a relations, composition and dependency of concepts are expressed by definitive relations. The annotation also supports narrative worlds inside narrative worlds, allowing for semantic entities to be annotated differently depending on the narrative world in question. E.g. a picture of someone holding a picture.

Since this kind of annotation focuses on narrative worlds, it does not address aspects of the image related to the representation of objects which are needed for a more complex image query: description of the poses of the figures in natural language terms (vs. the low-level contour-based shape descriptor values), the spatial configuration of the objects from a specific frame of reference, the stage of the event or the description of any motion involved in it, the type of setting or any background elements part of it, the way the participants are portrayed with respect to the viewer or with respect to other participants, etc. Spatial relations in MPEG-7 encode functions such as locatedIn or appearsIn, but they do not describe the spatial configuration of objects in detail. Likewise, the relations destinationOf, sourceOf, etc. provide information about the path of an object in motion, but do not characterize the motion itself. While expressing some of these points, like the way participants are oriented towards the viewer, can not be considered part of a purely semantic annotation task where the focus is primarily on concepts, they do affect the interpretation of the media
and cannot be captured by shape, texture or color visual descriptors alone. To what extent could a MPEG-7-based semantic annotation distinguish between an image of a person who has kicked a ball in a park, from an image of the same person in the same clothes, same location, and same time, but who is about to kick a ball towards the viewer of the image, making eye-contact with him while aiming?

2.4 Annotation Using Ontologies

Not all image annotation approaches use the semantic MPEG-7 vocabulary. Some works develop their own annotation schemes and integrate available Semantic Web ontologies in the annotation task, although low-level visual features are widely represented using the multimedia structure descriptors (Bloehdorn et al. 2005; Hollink 2006). The goal of multimedia annotation using ontologies is to create resources of multimedia content whose annotation has been mapped to well defined concepts in (multimedia) ontologies either available on the Web or designed for the annotation task. This line of work is particularly active in video annotation (Bertini et al. 2006).

Hollink et al. (2003); Hollink (2006)’s annotation and search tool provides a framework for annotation of art collections integrating domain thesauri (Architecture Thesaurus, Union List of Artist Names) and other ontologies (WordNet (Fellbaum 1998), Iconclass (Van der Waal 1974)). The content of the image is described in a series of statements of the form <agent, action, object, recipient>. The value for each of these fields is then mapped to the appropriate ontology. The setting is described with a frame structure <event, place, time>.\footnote{event in this case is understood as occasion, e.g. birthday, not as a TimeML event: “a cover term for situations that happen or occur” (Pustejovsky et al., 2003, pg. 29).} This kind of structured annotation resembles the MPEG-7 Structured Text Annotation element (Who, WhatAction, Where, When, Why, How).
Bloehdorn et al. (2005), on the other hand, present a software environment that allows the user to map low-level MPEG-7 visual descriptors to ontologies in the Semantic Web (e.g. DOLCE) in order to build ontologies that include prototypical instances of the high-level concepts and a formal specification of its low-level visual features, i.e., a knowledge base where concepts are associated with instances regarded as ‘prototypical’ in terms of their visual features. Some topological and directional relations among objects are also supported, namely, *adjacency*, *below*, and *inclusion*.\(^6\)

**Problems**

Hollink et al. (2003)’s approach seems to be halfway between keyword annotation and MPEG-7-based semantic annotation: although it encodes semantic relations between events and participants and structured setting and time information, the annotation is still limited in terms of a description of participant’s properties and representation, definitive or spatial relations among each other or motion description. In the context of art history, however, it provides a framework allowing for a rich annotation than complements the frames recording meta-data about title, creator, materials, location, subject keywords, etc.

Bloehdorn et al. (2005)’s knowledge base of prototypical instances is based upon the fact that linking low-level visual features with canonical instances of an event would presumably lead to improved action recognition systems. As MPEG-7-based approaches, however, it does not cover details about the representation of objects, their posture, orientation, or motion.\(^7\)

---

\(^6\)DOLCE ontology: [http://www.loa.istc.cnr.it/old/DOLCE.html](http://www.loa.istc.cnr.it/old/DOLCE.html)

\(^7\)The Motion Descriptor is used in the annotation of video sequences to capture the motion of the camera and the activity level or pace of motion of the scene (Sikora 2001), not the information related to the motion an entity undergoes in an event.
2.5 Description Generation

Captions in the form of sentences usually describe image content by identifying an event, its location, its participants and their properties (Figure 2.5) and sometimes their spatial configuration with respect to the *image structure* (Pustejovsky and Yocum 2014, section 2.6.2), as in The person in the center is wearing a helmet and riding a dirt bike. Efforts in image annotation and computer vision have promoted the development of databases of images annotated with sentence descriptions (e.g. UIUC Pascal Sentence Dataset (Rashtchian et al. 2010), the ImageCLEF IAPR TC-12 Benchmark Dataset) that serve as training for caption generation models either by directly extracting information from the captions (Luo et al. 2009; Elliott and Keller 2011, 2013; Gupta 2013) or by annotating those captions using annotation schemes like ISO-Space (Pustejovsky et al. (2013); Pustejovsky and Yocum (2014)) first, which, to my knowledge, has not been done yet. This section introduces some of the approaches in automatic content description, but it is important to mention that the focus here is on object recognition, image similarity, and natural language generation rather than on the annotation of semantic entities, although visual dependency graphs (see below) do provide a structured representation of the spatial configuration of objects in the image.

Elliott and Keller (2011) create a treebank of images of activities annotated with two sentences describing the event and the setting, which are later used as guide for object annotation with the LabelMe tool (Russell et al. 2008). This tool allows users to draw polygons around the objects and enter a label for them (e.g. tree). Object annotations are in turn used to generate a visual dependency representation (VDR) of the content (Figure 2.6), going back to dependency grammar in linguistics. Elliott and Keller (2013) draw upon this idea of VDRs, and use them, along with a corpus and sentence captions, to learn a model that automatically constructs sentence descriptions for the images of the PASCAL Visual

---

8ImageCLEF Dataset: [http://www.imageclef.org/photodata](http://www.imageclef.org/photodata)
An updated and extensive list of different image databases used for computer vision tasks is available through Prof. Robert Fisher’s homepage at the University of Edinburgh: [http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm](http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm)
CHAPTER 2. IMAGE ANNOTATION

object classification dataset matching a set of language templates. Farhadi et al. (2010) also apply VDRs and corpus statistics, but include object detectors to map a set of images and a set of sentences to an intermediate meaning space where instances are of the form \((\text{object}, \text{action}, \text{scene})\). This allows for both automatic image description (image \(\rightarrow\) meaning space \(\rightarrow\) sentence space) as well as text illustration (sentence space \(\rightarrow\) meaning space \(\rightarrow\) image space). Gupta (2013), on the other hand, exploits dependency relations in the captions and image similarity to provide a description for a given image by combining triples of the form \(((\text{attribute}_1, \text{object}_1), \text{verb}), (\text{verb}, \text{prep}, (\text{attribute}_2, \text{object}_2)), (\text{object}_1, \text{prep}, \text{object}_2))\).

Figure 2.5: Sample images and annotations from the UIUC Pascal Sentence Dataset (Rashtchian et al. (2010))

Figure 2.6: Visual Dependency Representation (VDR) of an image of a man sitting on a chair with his laptop (Elliott and Keller 2014)

To my knowledge there is not any dataset of images with captions previously annotated with schemes that capture spatial information in text such as SpatialML (Mani et al. 2010),

---

ISO-Space (Pustejovsky et al. 2013; Pustejovsky and Yocum 2014), or Kordjamshidi et al. 2010’s scheme for Spatial Role Labeling. Likewise, there is not machine learning research done in automatic caption generation using captions annotated with any other annotation scheme for natural language text. The notions the three schemes mentioned above address, however, are worth considering into an image annotation task and are therefore presented in the next section.

2.6 Annotation of Spatial Information in Natural Language Text

SpatialML (Mani et al. 2010), ISO-Space (Pustejovsky et al. 2013; Pustejovsky and Yocum 2014) and Kordjamshidi et al. 2010’s scheme for Spatial Role Labeling capture spatial information encoded in text in different ways. Two of them have originally been designed to annotate news text (SpatialML, ISO-Space), and SpatialML has also been applied to a medical corpus (ProMed) and the US. Immigration and Customs Corpus (ICE). Pustejovsky and Yocum (2014) suggest a version of ISO-Space adapted to image annotation. This section provides a brief overview of these annotation schemas, many of whose elements and covered concepts have been included in the annotation model presented in chapter 3.

2.6.1 SpatialML

SpatialML (Mani et al. 2010) marks up locations mentioned in a text and the relations among them, grounding them when possible to their geographic coordinates and mapping them to geographical resources like databases and gazetteers. Locations are tagged by the place tag and can be of different type, including STATE, COUNTRY, ROAD, VEHICLE, among others. The tag signal applies to expressions that trigger a relation between two places. Relations can be of two types: rlinks capture information about the orientation
CHAPTER 2. IMAGE ANNOTATION

```xml
a <PLACE id="1" form="NOM"> building </PLACE>
<SIGNAL id = "2" type = "DISTANCE">5 miles</SIGNAL>
<SIGNAL id = "3" type = "DIRECTION">east</SIGNAL> of
<PLACE id = "4" country = "TW" form = "NAM" latLong = "22?370N
120?210E"> Fengshan</PLACE>
<RLINK id = "5" source = "4" destination = "1" distance = "2"
direction = "E" frame = "VIEWER" signals = "2 3"/>
```

Figure 2.7: SpatialML annotation example (Mani et al. 2010)

(behind, above, east of, in front of, etc.) and distance between two locations, while LINKS the
topological relation between the regions of both locations by means of the Region Connection
Calculus (Randell, D. A. et al. 1992). The frame of reference is also encoded in orientation
relations through PLACE’s frame attribute values EXTRINSINC (absolute), INTRINSIC or
VIEWER (relative). In terms of figure and ground, the absolute frame assumes a coordinate
system and cardinal orientations that originate in the ground object (e.g. north, east); the
intrinsic frame involves the faces of the ground object (e.g. front, back), and the relative
frame involves the point of view of a third element, the viewer (Levinson 2003, as cited in
Mani and Pustejovsky (2012)). Locations themselves can be defined through orientation
relations (e.g. Southern U.S) by using the mod attribute of the PLACE tag. The annotation
can be automatically integrated with the Geography Markup Language (GML) defined by
the Open Geospatial Consortium (OGC) and to KML (Google Earth’s Keyhole Markup
Language (KML).  

10 The RCC defines eight different relations between regions (here, the regions occupied by the locations)
that express mereotopological information:

a. Disconnected (DC): A and B do not touch each other.
b. Externally Connected (EC): A and B touch each other at their boundaries.
c. Partial Overlap (PO): A and B overlap each other in Euclidean space.
d. Equal (EQ): A and B occupy the exact same Euclidean space.
e. Tangential Proper Part (TPP): A is inside B and touches the boundary of B.
f. Nontangential Proper Part (NTPP): A is inside B and does not touch the boundary of B.
g. Tangential Proper Part Inverse (TPPi): B is inside A and touches the boundary of A.
h. Nontangential Proper Part Inverse (NTPPi): B is inside A and does not touch the boundary of A.

(Randell, D. A. et al. 1992, as cited in Mani and Pustejovsky (2012))

11 GML: http://www.opengeospatial.org/standards/gml
KML: https://developers.google.com/kml/documentation/
2.6.2 ISO-Space

ISO-Space (Pustejovsky et al. 2013) builds upon SpatialML and is designed to represent spatial and temporal relations expressed in natural language texts. It captures topological relations between two regions as well as orientation relations and the different frames of reference. Furthermore, it encodes metric properties of elements, matric values between regions and objects and the motion of objects and its characteristics. The tag place, based on SpatialML place tag, marks up geographical (lakes, rivers, etc.) and administrative entities (cities, towns, etc.), that is, elements with geographical coordinates that can be mapped to Google Maps, if available. place attributes draw upon SpatialML place’s ones, with some modifications.\footnote{place tags encode whether the place denoted by the extent is the document creation location through the attribute dcl. The mod attribute is not related to orientation, in contrast to its SpatialML counterpart, but with any entity modifier that provides spatial information.} A path is a location that links two regions and can be traversed. Its set of attributes includes those from the place tag and also attributes capturing the origin and end of the path. Elements that do not denote locations but are used as a region in space in language are represented through the spatial named entity (spatial-ne), event, and motion tags (events that indicate a change of location). motion tags identify in their attributes the kind of motion verb (manner-path distinction (Talmy 1985)), and to which class of motion verbs it belongs.\footnote{Events are previously detected and annotated with TimeML, following a layered approach (Pustejovsky et al. 2013). The classes of motion verbs are taken from Pustejovsky and Moszkowicz (2008), which was based on the motion classes in Muller (1998): MOVE, MOVE _EXTERNAL, MOVE _INTERNAL, LEAVE, REACH, DETACH, HIT, FOLLOW, DEViate, STAY.} A spatial signal provides information to a link tag, and can be topological or directional. Links encode different information depending on the type: qlinks (example (1-a)) express topological relations between elements using RCC + IN; olinks encode orientation relations (near, above, front, left, north, etc.) between two entities in terms of a frame of reference (ABSOLUTE, INTRINSIC, RELATIVE) and a reference point (cardinal direction, ground entity, viewer entity); movelinks (example (1-b)) are triggered by motion events and record information about the entity in motion and the path traversed; lastly, metric links (mlink) define the distance between two regions or
the dimensional properties of an entity.

(1) a. \[ \text{The book}_{\text{sne}1} \] is \[ \text{on}_{\text{s1}} \] \[ \text{the table}_{\text{sne}2} \].

\[
\begin{align*}
spatial\_signal(s1, \text{cluster=on-1, semantic\_type=topological, directional})
\end{align*}
\]

qslink(qsl1, figure=sne1, ground=sne2, trigger=s1, relType=EC)

b. \[ \text{John}_{\text{sne}1} \] \[ \text{walked}_{\text{m1}} \] from \[ \text{Boston}_{\text{pl}1} \] to \[ \text{Cambridge}_{\text{pl}2} \].

\[
\begin{align*}
motion(id=m1, \text{motion\_type=MANNER, motion\_class=MOVE})
movelink(mv1, trigger=m1, source=pl1, goal=pl2, mover=sne1, goal\_reached=TRUE)
\end{align*}
\]

(Pustejovsky et al. 2013)

In their adaptation of ISO-Space to image annotation, Pustejovsky and Yocum (2014) point out that captions can refer either to structural regions of the image (the corner, the middle, etc.) or to content-dependent features of the objects in the picture. They draw the distinction between content configuration (where an object is situated relative to another object in the image) and structural configuration (where an object is located within the structure of the image). This distinction is reflected in the attribute domain of the extent tags, which takes as value STRUCTURE or CONTENT. The distinction allows for the document creation location to have two different values, one, Image Structure Location, referring to the image as an object (View of New York City at night), and the other one, Image Content Location, related to the content of the image (New York City). OLINKS are assumed to have a relative frame of reference with the viewer as point of reference, unless otherwise stated. Part of the paper shows how this distinction between structure and content in the annotated entities can help to infer new relations between objects by means of a compositionality table.
2.6.3 Spatial Role Labeling

Kordjamshidi et al. (2010) turn to Semantic Role Labeling (Marquez et al., 2008) and Holistic Spatial Semantic (HSS) (Zlatev, 2003; Zlatev, 2007) to design an annotation scheme that marks up spatial expressions (spatial indicators) and classifies their arguments in terms of their spatial roles. Each token can have different roles if it participates in different relations. The main unit of analysis is the sentence, not a specific linguistic category (e.g. prepositions), hence ‘holistic’. The components of HSS theory are the following (Kordjamshidi et al., 2010, pg. 414): trajector (the entity whose location or motion is of relevance), landmark (the reference entity in relation to which the trajector is determined), region (a region of space defined in relation to the landmark), a path (characterization of a path or a complex landmark in relation to a region defined by the landmark in terms of its beginning, middle and end), direction (direction along the axes provided by the different frames of reference), and frame of reference. Spatial indicators define the type of spatial relation they trigger, and express region, direction or distance. If the type of the spatial indicator is region, the relation expresses topological relationships through a mapping to RCC-8. A direction type triggers directional relationships defined in terms of a frame of reference, and distance spatial indicators express qualitative or quantitative distances.
Chapter 3

A Model for the Semantic Annotation of Images

3.1 Overview

Tagging approaches do not provide richly structured information about the content of the image beyond the classification of tags in who, what, where, when categories presented in some of them. Spatial or semantic relations among the objects in the image are not expressed. The use of ontologies enables mappings of image labels to available ontologies in the Web. Though some efforts include a richer annotation frame, the focus is mainly on the integration of online resources in the annotation, the association of low-level features to concepts from different ontologies, and the creation of multimedia ontologies that store both instances of multimedia data and the concepts to which they are mapped. This would improve object and event recognition systems developed in the computer vision community. On the other hand, MPEG-7 descriptors capture the narrative world of an image and semantic relations between participants of an event to some extent, but they do not encode aspects related to spatial information, the motion of an object or details about the location that are thoroughly
accounted for in text annotation schemes. In an attempt at image annotation there are several points that an scheme could considerate, three of which have already been addressed in the projects mentioned in the last chapter and are listed here:

1. Record the event presented in the image, its participants, and their semantic roles.
2. Capture the spatial configuration of objects and characterize their motion, if the event involves any.
3. Provide information about the setting and location of the situation depicted in the image.

Additionally, the following aspects regarding the different elements of an image should be accounted for as well:

**Objects**

1. The representation of objects from the viewer’s perspective, including their orientation and perceivable extent.
2. Categorization of objects using a richer set of classes that includes *Person* and *Object* (MPEG-based), among others.
3. Physical appearance of objects (and emotional attitude, if applicable).

**Events**

1. Stage of the main event captured in the image and subevents involved in it (e.g. *sitting*, *holding*, etc.)
2. Characterization of the motion expressed by an event in terms of the object in motion, other figures, and the viewer. In an image of two people running in different directions from the viewer’s perspective, the annotation would record the value associated with the motion of each figure: leftwards and rightwards.
3. Identify the different events an image might denote, in case it is ambiguous.
4. Classify the image instance as ‘(non-)prototypical’ given the main event and its visual features.
CHAPTER 3. A MODEL FOR THE SEMANTIC ANNOTATION OF IMAGES

Setting and Time

1. Location types. This is included in both SpatialML and ISO-Space, and can be adopted in image annotation as well.

2. Background elements and events. Tagging sometimes involves references to elements in the background, but capturing which elements and events occur in the background and which ones are part of the foreground would be significant to event recognition and automatic content description.

Use cases

Querying

An XML schema capturing the above mentioned information would allow for an accurate structured retrieval, and different kinds of queries could be taken:

a. Queries by event or semantic role

   i. Queries with a specific semantic role: e.g. a query of pictures of events where the instrument is a knife should yield pictures of people cutting, spreading butter, engraving, carving, or even stabbing; queries in which the event read must involve a recipient would give photos of people reading to someone and exclude those in which there is only a person reading, etc.

   ii. Queries that specify a stage of an event: e.g. pictures of people about to jump from a springboard vs. pictures of athletes already in the air or in the swimming pool.

   iii. Queries of an event that is a subevent of another event: e.g. queries of pictures where event = sit should return pictures of people watching television, studying, reading, eating, etc. as long as the figures in those images are indeed sitting.

   iv. Queries that refer to the direction of an object in motion: e.g. pictures of people going in reverse.

   v. Queries in which the manner an event is carried out is not the prototypical one (e.g. sleeping standing, in unusual poses) or in which the images do not match our canonical visualization of the event. For instance, not prototypical results for the events watching TV, walking the dog or driving would be pictures of someone watching TV where the TV is not visible, images of someone walking the dog and wearing a costume, or images of people driving an uncommon car, like a 1920’s replica, respectively.

b. Queries by the representation of figures and camera angle

   i. Queries by angle view. E.g. pictures of someone climbing taken from a high angle.
ii. Queries by the viewer’s point of view: the perceivable extent of the figures or the direction of their motion relative to the viewer. E.g. images of people running in which the torso is not visible, pictures of someone giving a speech from a rear view, etc.

iii. Queries in terms of the appearance of objects: their physical and emotional characteristics, their outfit, etc.

Computer Vision

A database of images with structured information describing their content in natural language text could serve to learn mappings between annotated extents and its corresponding low-level features in computer vision applications. Action recognition in pictures needs object recognition to assist it. Capturing the way objects are represented, along with a characterization of their visual appearance, could be of use in the object recognition task. Gupta and Davis (2007) show that recognizing the kind of objects, the type of action in the image (in terms of poses of figures) and the object’s reaction to that action, in turn, improves the overall event recognition. Li and Fei-Fei (2007) also prove that object and setting recognition are significant clues to event recognition. Thus, recording subevents that require a figure’s specific posture (e.g. holding), the type of objects and their perceivable extent in the image, as well as details about elements of the setting, would help the system to make better predictions.

A rich annotation of images could also be valuable for the automatic content description task (caption generation). In their study to discover salient features of an image that are usually the object of descriptions, Berg et al. (2012) state that the type of object, type of scene, the object’s attributes and the image context are significant features to which annotators draw their attention when writing a description of the image. Striking features,

---

1 Siskind (1999) shows that this is not necessarily the case of motion events in movie clips. In his approach to extract significant visual features to learn a model to recognize motion events in video, objects regions are reduced to ellipses: tracking the movement of the ellipses and the mereotopological relations through time gives significant clues to the subevents recorded in the different clips, which ultimately leads to recognition of the main event.
like *pink* in *pink elephant*, objects not commonly found in specific settings (e.g. *a bicycle in a kitchen*), as well as indoor settings (in contrast to outdoor ones), are likely to be described in captions. A structured annotation of these aspects would thus contribute to this effort.

### 3.2 Model

In this section a model for the semantic annotation of images is proposed. The conceptual schema provides a brief introduction to the information it covers, its elements and the relations among them.

#### 3.2.1 Conceptual Schema

Every image has an element `IMAGE` that captures general information about the image in terms of its type and camera angle (see Figure 3.2.1). Since the model is designed to annotate images depicting events, in every image there is at least one `EVENT` element. If the event is ambiguous, alternative events can be included and related to each other with an `EXLINK` tag (mutually exclusive). The events in the image always happen in one specific location or setting at a specific time. The elements marked as `FIGURES` in the annotation can be related to an event by a `ROLELINK`, which expresses the semantic role they play in the event. If the event involves motion, there is a link that relates the event and the moving object and which captures information about the movement and the motion path. This movement is further characterized with a `DIRECTIONLINK`, that encodes the direction of the figure in motion according to different frames of reference. Figures, in turn, can be represented in the image or be omitted (`OFIGS`) if they do not appear in the image but are assumed by the viewer in order to interpret the event(s). Their orientation with respect to the viewer along a vertical axis (e.g. front view, back view), their perceivable extent (head, waist-up, bottom-part, etc.), and their visual appearance are recorded as well. Furthermore, `FIGURES`
can hold, face and look at another figures. The distinction between facing relations and gazing relations relies on the fact that a FIGURE does not need to face another in order to look at it and vice versa.

3.2.2 Elements

3.2.2.1 Image

The IMAGE elements encodes information about the image as a whole. The values of its attribute type come partly from Google Advance Image Search types (Face, Photo, Clipart,
CHAPTER 3. A MODEL FOR THE SEMANTIC ANNOTATION OF IMAGES

Line Drawing, Animated), with the addition of painting and 3D_Shot (for images from videogames). The attribute angle refers the camera angle from which the photo was taken or to the perspective of a painting or drawing: a photo is taken from below (low), from above (high) or from an eye-level position (neutral). The camera angle affects the viewer’s interpretation of the properties of figures in the image: a person looks taller and bigger from a low camera angle, smaller from a high one. The attribute prototypical refers to the event represented on the foreground and goes back to Bloehdorn et al. (2005)’s prototyping approach (2.4).

\[
\begin{align*}
\text{iid} & \::= \text{ID} \\
\{\text{iid} & \::= \text{ImageID} \\
\text{ImageID} & \::= i<\text{integer}>\}
\end{align*}
\]

\[
\begin{align*}
\text{type} & \::= \text{‘PHOTO’ | ‘PAINTING’ | ‘DRAWING’ | ‘CLIPART\_CARTOON’} \\
& \text{ | ‘3D\_SHOT’ | ‘OTHER’ \{default is ‘PHOTO’\}}
\end{align*}
\]

\[
\begin{align*}
\text{angle} & \::= \text{‘NEUTRAL’ | ‘HIGH’ | ‘LOW’ | ‘OTHER’} \\
\text{prototypicalInstance} & \::= \text{‘yes’ | ‘no’}
\end{align*}
\]

3.2.2.2 Figure

The tag figure refers to entities in the image that either take part in an event or, if not, they participate in a HOLDINGLINK relationship with another entity figure that does (in a HOLDINGLINK, a figure holds another figure). FIGURES are usually in the foreground, unless there is some activity represented in the background as well, and are the most salient elements of the image. They usually correspond to objects or living beings, although geospatial entities and facilities can be FIGURES as well. Entities that are FIGURES occupy a specific region of the image and have a perceivable contour, in contrast to elements that conform the whole SETTING of the image (see the SETTING tag): In an photo of a train arriving at a station taken from inside the station, there is not a figure station; the station
is the whole setting of the image and does not have a distinguishable shape in the picture. However, if the shot is taken from outside and the building of the station is visible, this would be expressed with a FIGURE tag. Having a distinguishable contour is important here because FIGURE elements could be mapped to their low-level features such as shape (contour-based) for visual recognition systems. A FIGURE is involved in an EVENT if takes one of the values for semantic role in the ROLELINK tag between a FIGURE and an EVENT. Entities not involved in an FIGURE or a HOLDINGLINK are not tagged as FIGURES, but mentioned in another part of the specification. The idea behind this is to only capture the represented activities and the properties of their participants, not the characteristics of the rest of the elements in the image. Although FIGURES seem to be the same as objects in Object Recognition Task, there is a small difference between those concepts: in an Object Recognition Tasks, a tree in the background is an object, but it would not be recorded as a FIGURE here if it is not related to an EVENT or is an argument of a HOLDINGLINK. The term figure is related to the concept of figure in Talmy (1975), whose distinction between figure and ground is captured in ISO-Space relation tags (2.6.2). In the example of a pen falling off a table, the pen is analyzed as figure in Talmy (1975) and would be tagged as FIGURE in this schema, but the table, the ground according to Talmy, would be captured here as FIGURE as well and referenced in the attribute source in a MOTIONLINK.

The attribute type refers to the category to which the figure belongs. Part of its values (Facility, Geographical Entity) come from ACE Entity Types (FAC, GPE, LOC, ORG, PER), Object and Person come from the ontology of semantic types from the MPEG-7 semantic descriptors, and Vehicle is one of the types of the tag PLACE in SpatialML (2.6.1). The view attribute encodes the way a FIGURE is represented with respect to the viewer in terms of its position along a vertical axis and the values are drawn from portrait painting techniques. How much of the figure is perceivable is expressed by the extent attribute. The values top_part, bottom_part, and inside apply only to inanimate entities, whereas waist_up,

waist_down, bust and and knees_feet to people and animals. The single_part value covers those cases in which only one specific portion of the figure is visible, which is then indicated as value of the attribute singlePart. The attribute state records relevant properties of the figures that last for a period of time, e.g. an open door, a lamp turned on, a pregnant woman, a hot cup of coffee, or a building under construction. The emotional states of people and animals, if evident in the picture, are captured by the attitude attribute. Adjectives such as angry, happy or confused are some of the possible values it can take. The physAspect attribute describes any visual property of the figures that would match information captured by low-level features such as color, texture or shape. These usually translate to a brief mention about the outfit of a person, her hair color, or her age; about the color, size or shape of an object; or the size, breed or subspecies of an animal, etc.

figid ::= ID
{figid ::= FigureID
  FigureID ::= fig<integer>}
type ::= ‘PERSON’ | ‘VEHICLE’ | ‘FACILITY’ | ‘OBJECT’ |
  ‘ANIMAL’ | ‘PLANT’ | ‘GE’ | ‘OTHER’
view ::= ‘FRONT’ | ‘3/4’ | ‘PROFILE_LATERAL’ | ‘BACK’ | ‘OTHER’
extent ::= ‘WHOLE’ | ‘TOP_PART’ | ‘BOTTOM_PART’ | ‘INSIDE’ |
  ‘LATERAL’ | ‘WAIST_UP’ | ‘WAIST_DOWN’ | ‘BUST’ |
  ‘KNEES_FEET’ | ‘SINGLE_PART’ | ‘OTHER’
singlePart ::= CDATA
state ::= CDATA
attitude ::= CDATA
physAspect ::= CDATA
comment ::= CDATA
CHAPTER 3. A MODEL FOR THE SEMANTIC ANNOTATION OF IMAGES

3.2.2.3 Ofigure

The OFIGURE tag captures omitted figures. It is used to include those participants of an event that are not represented in the image but are assumed by the viewer in order to interpret the content. The viewer herself can function as an omitted figure in images in which e.g. someone is handing a book to the camera or pointing at it. In those cases, the word ‘VIEWER’ is entered as value of the comment attribute. Further examples of events with omitted figures are e.g. a picture of someone giving a speech in which the audience is not represented, or a close-up shot of someone writing a letter where only a pen and a notebook appear in the photo. In this last example, the OFIG plays the role of the agent, who presumably is holding the pen (instrument). The attributes of the tag FIGURE do not apply to OFIGs, since they do not have any visual property, but OFIGs take part in links the same way as figures do.

\[
\text{ofigid ::= ID}
\]

{\text{ofigid ::= OFigureID}

\text{OFigureID ::= ofig<integer>}}

\text{comment ::= = CDATA}

3.2.2.4 Event

The tag EVENT encodes activities or events represented in the image, but not states, which are only expressed via the state attribute of FIGURES. The term EVENT for this tag is inspired by the tag of the same name in TimeML (Pustejovsky et al. 2003) and the tag ID has the same form. EVENTS roughly equate to SemanticBase elements of type EventType in the MPEG-7 semantic description terminology (2.3), to the WhatAction field of the MPEG-7 StructuredText frame, or to action in both Hollink et al. (2003) and Farhadi et al. (2010). As in these approaches, the text is entered in infinitive form: jog, shake hands, drink, etc.
CHAPTER 3. A MODEL FOR THE SEMANTIC ANNOTATION OF IMAGES

The stage attribute captures the phase of the event represented in the image. Searching in Google for images of propellent motion verbs (Miller 1972), like throw, push, or kick, usually gives mixed results of images of the event when it is about to happen, while it is happening, or moments after it happened. Since the event is given in its infinite form and aspect is not encoded, stage accounts for this distinction. The type attribute identifies the event as a subevent (an event that is part of another event) or as a main event (an event which is not a subevent). Events related to the posture of the figures, like sitting, lying down, smiling are common subevents of main events (e.g. reading sitting on a sofa, sleeping lying down in bed, etc.). If a figure is involved in more than one event, determining which one is the main event and which one the subevent might be a subjective decision in some images. In an image depicting someone talking on the phone while driving, drive could be marked up as an event of type MAIN, while talk as its subevent. Another example of a subevent would be listen to music in a picture of someone waiting at a bus stop wearing headphones, where the main event is wait. In contrast, it is trickier to make the distinction in the case of a picture of someone jumping and taking a photo at the same time. The subevents of a main event are listed in the attribute subevent, which takes EventIDs as values. The attribute manner refers to information related to the way the event occurs if that information is not already encoded in the event or in any of its subevents. In motion verbs, this relates to the manner and path distinction in verbs addressed in Talmy (1985) and encoded in the motion-type attribute of motion tags in ISO-Space. Here, manner just expresses any kind of modal information that is not considered a subevent of the main event (walking with hands in the pockets, exercising vigorously, driving fast, crawl swimming). holdingLink (see below) refers to HOLDINGLINK relations between two figures. The attribute representsConcept comes from the Concept abstract descriptor schema of MPEG-7, where an event shaking hands is linked to a Concept Comradeship.³

³A list of keywords that would be useful for marketing, advertisement or management can be entered here. E.g. progress, improvement, ambition as concepts for an image of someone climbing a mountain, or patience for an image of someone fishing. This concerns images known as motivational posters in the Internet community.
3.2.2.5 Setting

The tag setting expresses information about the location in which the events occur and the elements appearing in the background. It refers to general locations such as park, kitchen, or beach. Sometimes an event happens in a figure part of the setting, such as climbing a tree in a park, reading a book on a sofa or talking on the phone inside a car in a parking lot. Tree, sofa, and car are figures because they occupy a specific region on the image, have a distinguishable contour, and do not constitute the whole setting of the picture.\textsuperscript{4} In those cases, the figures are assigned the role PLACE in the rolelink that links event and figures and they can be involved in other events as well via other rolelinks. These figures are included as part of the setting by the attribute figureID. The general type of scene is captured by the attribute scene, that mainly distinguishes between outdoor and indoor

\textsuperscript{4}Likewise, in a image of two monkeys eating fruit in a canopy of a tree, where the canopy conforms the setting of the whole image, tree would not be a figure, it would be the setting.
CHAPTER 3. A MODEL FOR THE SEMANTIC ANNOTATION OF IMAGES

scenes, with solid_background being used for images with blank or blurred background. The specific type of scene attribute builds upon the values for the attribute medium in motion events (see motionlink tag) and includes values from location type from ISO-Space. Significant elements in the background of an image which are not figures are listed as values for backgroundElements, while figures in the background are given in backgroundFigures. If the exact geographical location is known (e.g. Madison Square), it is encoded by the geoInformation attribute.

\[
\begin{align*}
\text{sid} & ::= \text{ID} \\
\{\text{sid} & ::= \text{SettingID} \\
\text{SettingID} & ::= s<\text{integer}>\} \\
\text{figureID} & ::= \text{IDREF} \\
\{\text{figureID} & ::= \text{FigureID}\} \\
\text{scene} & ::= '\text{OUTDOORS}' | '\text{INDOORS}'| '\text{SOLIDBACKGROUND}'| '\text{OTHER}' \\
\text{type} & ::= '\text{STREET}' | '\text{FACILITY}' | '\text{LANDSCAPE}' | '\text{ROAD}' | '\text{WATER}' | '\text{AIR}' | '\text{SOLIDBACKGROUND}' | '\text{OTHER}' \\
\text{backgroundElements} & ::= \text{CDATA} \\
\text{backgroundFigures} & ::= \text{IDREF} \\
\{\text{backgroundFigures} & ::= \text{FigureID}\} \\
\text{geoInformation} & ::= \text{CDATA}
\end{align*}
\]

3.2.2.6 Time

The tag time records the time at which the events represented in the image happened (if it is a photo, this is the same as when the photo was taken).\(^5\) time only has one attribute,

\(^5\)In paintings this is more difficult. There is the time of the production of the work, which is part of the meta-data of the painting along with information about painter, format, etc. and the time of the events, which is what the time tag attempts to capture. However, events might become common subject matters of art (e.g. Jesus Christ’s Crucifixion, Alexander The Great Battles, etc.) and the representation of the figures does not always match the period of the event. This is related to iconography and outside the scope of this thesis. See the Iconclass ontology for more information about this: http://www.iconclass.nl/home.
type, which encodes general information about the time implied by the background of the image. The id form of this tag goes back to TIMEX id attribute in TimeML: $t0$, $t1$, $t2$, etc.

\[
\text{tid ::= ID} \\
\{\text{tid ::= TimeID} \\
\text{TimeID ::= t<integer>}\}
\]

\[
\text{type ::= ‘DAYTIME’ | ‘NIGHTTIME’ | ‘SUNRISE_SUNSET’ |}
\text{‘SOLID_BACKGOUND’ | ‘OTHER’} \\
\text{comment ::= CDATA}
\]

### 3.2.2.7 Rolelink

ROLELINKS take a FIGURE and an EVENT as arguments to express the semantic role the FIGURE plays in that event. The values of the role attribute are mostly drawn from VerbNet list of semantic roles (Kipper-Schuler et al. 2006 as cited in Palmer 2014), but do not include all of them:

**Agent, Co-Agent** applies mostly to FIGUREs that are volitional agents, usually people or animate subjects, but also applies to internally controlled subjects like machines.

**Patient** refers to FIGUREs that have been affected by the event, e.g. by a change of state.

As in VerbNet, also applied to FIGUREs involved in events of combining, attaching, separating, etc.

**Theme** FIGUREs that undergo a change of location.

**Experiencer** FIGUREs that are experiencing something. Used in images representing perception and psychological events, or events involving a physical stimulus.

**Stimulus** FIGUREs that trigger a response from a FIGURE experiencer.
CHAPTER 3. A MODEL FOR THE SEMANTIC ANNOTATION OF IMAGES

Medium This is drawn from FrameNet (Baker et al. 1998) *obsuring_medium* role from the frame PERCEPTION_ACTIVE, where it refers to the medium through which a phenomenon makes itself perceivable to the Perceiver_agentive – here, Experiencer –. E.g. *Someone eavesdropping through a wall*. Here it also refers to the direction role of that same frame, which is used in the opposite direction, e.g. *the detective listened through the wall*, and expresses how the attention of the perceiver is directed.\(^6\)

Recipient Covers both Recipient and Beneficiary roles of VerbNet. FIGURES that are the target of a transfer in events of change of possession, communication, and events involving the body, as well as FIGURES that benefit from an action (e.g. preparing).

Source and Goal FIGURES that are either the source or the destination or a movement.

Instrument FIGURES that are used by a FIGURE agent for some purpose.\(^7\)

Place This stands for location in VerbNet. Used to represent spatial configurations (see SETTING tag).

\begin{verbatim}
rlid ::= ID
{rlid ::= RoleLinkID}
RoleLinkID ::= rl<integer>}
figureID ::= IDREF
{figureID ::= FigureID}
eventID ::= IDREF
{eventID ::= EventID}
role ::= ‘AGENT’ | ‘CO-AGENT’ | ‘PATIENT’ | ‘THEME’ |
‘EXPERIENCER’ | ‘MEDIUM’ | ‘STIMULUS’ | ‘RECIPIENT’ |
‘SOURCE’ | ‘GOAL’ | ‘INSTRUMENT’ | ‘PLACE’
\end{verbatim}

\(^6\)More information about this distinction available in https://framenet2.icsi.berkeley.edu/fnReports/data/frame/Perception_active.xml, last accessed on 07/20/2014.

\(^7\)This differs from VerbNet’s definition: used for objects (or forces) that come in contact with an object and cause some change in them (Palmer (2014)).
3.2.2.8 Holdinglink

FIGURES that hold other FIGURES are linked by means of a HOLDINGLINK. This is also captured in the MPEG-7 description, where a new SemanticBase of entity type event would be added with value hold, a holding event. Holding events are a subevent of numerous events (e.g. eating a hamburger, throwing a ball, brushing your teeth, etc.), and most images would need an EVENT hold to account for that. For this reason, holding events are represented in this schema by their own tag. However, since in a holding event there is always the FIGURE holder and the FIGURE held, this structure is captured by a link instead of by an element of type EVENT. In this way, EVENTS have an attribute that refers to the HOLDINGLINKS involved it in, thus capturing holding events that are part of a main event. The attribute manner functions the same way as in EVENT entities.

\[
\text{hlid ::= ID} \\
\{\text{hlid ::= HoldingLinkID} \\
\text{HoldingLinkID ::= hl<integer>}\} \\
\text{holderFigureID ::= IDREF} \\
\{\text{holderFigureID ::= FigureID}\} \\
\text{heldFigureID ::= IDREF} \\
\{\text{heldFigureID ::= FigureID}\} \\
\text{manner ::= CDATA}
\]

3.2.2.9 Motionlink

MOTIONLINKS go back to ISO-Space MOVELINKS and take a FIGURE that moves, a FIGURE that is the causer of the movement (which can be the same as the FIGURE that moves), and an eventID associated with that motion as arguments. If the source and destination of the path of the movement are visible in the image, their ids are also included as values.
for the source and destination attributes respectively. If there is a physical path along
which the movement occurs, like a highway, a sidewalk, or a river, it is entered in the field
physicalPath. The medium attribute goes back to Miller (1972)’s medium of travel (land,
water, air) and is also related to ISO-Space version 1.0 path type, BodyofWater, road, mts.
If none of these values is applicable, the value other is selected and the appropriate medium
is given in the otherMedium attribute. MOTIONLINKs refer to DIRLINKs, which capture the
direction of the movement according to different FIGUREs as point of reference. For images
in which both agent and theme move together along in the same direction (bicycle and rider,
dog and dog walker, cart and cart pusher, car and driver) the ids of both are entered in the
figureID. This also holds for the attribute figureID in DIRLINK in these cases.

```
mlid ::= ID
{mlid ::= MotionLinkID
MotionLinkID ::= ml<integer>}
figureID ::= IDREF
causerID ::= IDREF
{causerID ::= FigureID}
eventID ::= IDREF
directionLinkID::= IDREF
{directionLinkID ::= DirectionLinkID
DirectionLinkID::= dl<integer>}
physicalPath ::= CDATA
sourceID ::= IDREF
{sourceID ::= FigureID}
destinationID ::= IDREF
{destinationID ::= FigureID}
medium ::= ‘LAND’| ‘AIR’ | ‘WATER’ | ‘OTHER’
otherMedium ::= CDATA
```
3.2.2.10 Dirlink

Dirlinks express information about the direction of the motion involved in the event, and their attributes come from SpatialML RLInKS as well as from the attributes figure, frame_type and referencePt in ISO-Space OLInKS. Dirlinks record the orientation of the movement from the perspective of the object in motion (intrinsic frame of reference), relative the causer of the movement, and relative to the viewer of the image. The values TOWARDS_VIEWER and AWAY_FROM_VIEWER account for the 3D properties of the content, whereas LEFT, RIGHT, TOP, BOTTOM, TOP-RIGHT, TOP-LEFT, etc. refer to the different regions of the image. E.g. a rear view of an athlete making a long jump would have the value FORWARD for intrinsic and relative_to_causer attributes, but AWAY_FROM_VIEWER relative to the viewer.

\[
\text{dlid ::= ID} \\
\{\text{dlid ::= DirectionLinkID} \} \\
\text{DirectionLinkID ::= dl<integer>} \}
\]

\[
\text{motionLinkID ::= IDREF} \\
\{\text{motionLinkID ::= MotionLinkID}\}
\]

\[
\text{figureID ::= IDREF} \\
\text{intrinsic ::= 'UPWARDS'| 'DOWNWARDS'| 'FORWARDS'| 'BACKWARDS'| 'LEFTWARDS'| 'RIGHTWARDS'| 'NONE'} \\
\text{relativeToCauser ::= 'UPWARDS'| 'DOWNWARDS'| 'FORWARDS'| 'BACKWARDS'| 'LEFTWARDS'| 'RIGHTWARDS'| 'NONE'} \\
\text{relativeToViewer ::= 'TOWARDS_VIEWER'| 'AWAY_FROM_VIEWER'| 'LEFTWARDS'| 'RIGHTWARDS'| 'TOPWARDS'| 'DOWNWARDS'| 'TOP_RIGHT'| 'TOP_LEFT'| 'BOTTOM_RIGHT'| 'BOTTOM_LEFT'}
\]
3.2.2.11 Facelink

Zitnick and Parikh (2013) show in their approach to discover semantically important features in images that eye gaze is a significant semantic cue in scene understanding. Since there are images in which two people are facing each other, but there is not eye contact between them, this idea has been divided here into two different relations: facing relations and gazing relations. Similarly, a FIGURE may be looking at another FIGURE without facing it. These two kinds of relation resemble the HOLDINGLINK tag, in which an event is captured using a link, instead of an element EVENT, because the argument structure of the event is always the same and the event is represented in numerous images. FACELINKS take two FIGURES as arguments and encode the way they are oriented to one another. The facing FIGURE can only be of type person or animal, while the faced one of any type. Eye-contact between two FIGURES facing each other is encoded in the attribute gazingEachOther to prevent the annotator from having to additionally create two gazing relations (one in each direction) in case the value is yes.

\[
\text{flid ::= ID} \\
\{\text{flid ::= FaceLinkID} \\
\text{faceLinkID ::= fl\langle integer\rangle}\} \\
\text{facingFigureID ::= IDREF} \\
\{\text{facingFigureID ::= FigureID}\} \\
\text{facedFigureID ::= IDREF} \\
\{\text{facedFigureID ::= FigureID}\} \\
\text{gazingEachOther ::= ‘YES’ | ‘NO’ | ‘N/A’ | ‘UNKNOWN’} \\
\text{comment ::= CDATA}
\]
3.2.2.12 Gazelink

A GAZELINK captures the event in which a FIGURE is looking at another FIGURE. GAZELINKS are triggered in cases in which a FIGURE is looking at another one and there is not a FACELINK relation between them, or, if there is one, the value of gazingEachOther is either not applicable (N/A) or NO.

```
clid ::= ID
  {clid ::= GazeLinkID
gazeLinkID ::= gl<integer>}
gazerID ::= IDREF
  {gazerID ::= FigureID}
gazedFigureID ::= IDREF
  {gazedFigureID ::= FigureID}
comment ::= CDATA
```

3.2.2.13 Exlink

EXLINKS take as arguments at least two EVENTS and express the fact that they are mutually exclusive. Some images might be ambiguous in the event they represent: a plane landing or taking off, someone parking the car or maneuvering to leave the spot, closing or opening a book, etc. If the actual event is unknown, both options should be included in the image as independent events and linked by means of an EXLINK. If low-level features were to be extracted, including both options allows for an association of the same low level features to both types of events. The disambiguation between the two events will depend on the text surrounding the image on a web page, which affects our interpretation of the image at the time of reading.

```
exid ::= ID
```
\{\texttt{exid ::= ExLinkID} \\
\texttt{ExLinkID ::= ex\langle integer\rangle}\}\}
\{\texttt{eventsIDs ::= IDREF} \\
\texttt{eventsIDs ::= EventID}\}
\{\texttt{comment ::= CDATA}\}

### 3.2.3 Annotation Examples

The following annotation examples illustrate the use of the model. The images taken from Google Image Search are under the Creative Commons license\(^8\).

\(^8\)http://creativecommons.org/
Figure 3.2: Extracted from Google Image Search. Source: Flicker user Marco Arment, Title: Brainstorming)
Figure 3.3: Extracted from Google Image Search. Source: www.20minutos.es
Figure 3.4: Extracted from Rashtchian et al. (2010)
Conclusion

The current ease to create images gives rise to the need for their efficient access and retrieval. Text-based image retrieval is based on queries by keywords in image captions or in the text that surrounds the image, for instance, in a web page. However, numerous images lack a caption, and the text, if any, does not always match image content. Unannotated images in web pages without text could not possibly be accessed using this method. Complementary to text-based image retrieval, content-based retrieval methods approach this task through queries by the image’s low-level features such as texture, shape or color. Research in computer vision areas like automatic object recognition and event detection using low-level features are thus helping in the automatic annotation of image content in order to improve image retrieval and content filtering. Projects in automatic generation of textual descriptions work towards this goal as well, and their contributions play a significant role for visually impaired users. Semantic annotation approaches attempt to bridge this gap between low-level features and users’ high-level descriptions to allow for queries by semantic content, in contrast to queries by keyword match in the text of a web page and queries by low-level features.

Tagging approaches capture the essential content of an image (see 2.2). However, they do not provide a complexly structured annotation, and the categorization in who, what,
when, where classes included in some collaborative tagging approaches does not account for notions that are needed for more complex queries (e.g. stage of the event, description of motion, etc.). The MPEG-7 format, on the other hand, defines a standard to store information related to the image’s low-level features by means of its descriptors. Annotation approaches based on its Semantic Descriptors capture semantic entities and their roles in the event depicted in the image, but they do not encode information about the representation of figures with respect to the viewer or to other figures, or the motion involved in an event. Lastly, image annotation using ontologies makes use of available ontologies in the Web to link the annotation and the image’s low-level features to concepts defined in them in order to improve computer vision systems. Efforts in automatic caption generation have promoted the development of databases with images annotated with sentence captions and models to represent the content encoded in those descriptions have been proposed as well (see Visual Dependency Graphs, section 2.5). Although some automatic caption generation approaches capture the spatial configuration of objects using visual dependency graphs as part of the process, the focus here is to design a model not to represent the content of the image as richly as possible, but to allow for sentence generation to be performed accurately. Object detectors and image similarity play thus an important role in this task.

Text annotation schemes address certain notions that are worth capturing in an image annotation task, such as orientation (of figures in the image), types of location, and the motion involved in events. SpatialML (Mani et al. 2010, see 2.6.1) marks up locations denoted by natural language text expressions and encodes directional and topological relations among them. ISO-Space (Pustejovsky et al. 2013, see 2.6.2) builds on this and extends the annotation to expressions that are not locations but still denote a spatial region. It characterizes the motion involved in motion events, as well as topological, orientation and metric relations between entities. Kordjamshidi et al. (2010) (see 2.6.3) approach the task of spatial information annotation as one of semantic role labeling and capture the arguments of spatial expressions as well as topological, directional and distance relations between them.
Chapter 3 presented a model for the semantic annotation of images of activities and events that attempts to integrate some of the aspects covered by text annotation schemes to allow for a better image retrieval. It expresses information about the type of image (photo or painting, angle, etc.) as well as details about the time and setting. Figures are characterized in terms of their appearance, their orientation with regard to the viewer, and their perceivable extent. A specific link tag captures the different semantic roles figures play in the event(s) depicted in the image. Cases in which figures face or hold other figures are also addressed. Events are identified as main or sub-events, and their phase is also recorded. If an event involves motion, a relation expresses the object in motion, the causer of the movement and information about the path and medium of the movement. The direction of the motion is defined then in terms of the object in motion, the causer of the movement, and the viewer of the image. In cases in which an image is ambiguous and might denote more than one event, the different alternative events can also be listed in the annotation and marked as mutually exclusive.

This model is, however, merely a first step. An application that let the user load images and draw polygons around the figures in order to annotate them with the suitable tag, in a similar way to LabelMe (Russell et al. 2008), would be extremely helpful for the annotation. Regarding the annotation process itself, the Model-Annnotate-Model-Annnotate Cycle (Pustejovsky and Stubbs 2012) should be performed with a group of different annotators in order to spot inconsistencies and come to a thoroughly tested and revised annotation model. On the other hand, spatial information, the focus of the text annotation schemes mentioned here, is not fully captured: spatial configuration of objects in relation to other objects is not addressed (beyond basic holding relations) and topological relations are not captured either. Since text annotation approaches are already moving towards the annotation of spatial information in image captions (Pustejovsky and Yocum 2014, see 2.6.2), this model was presented as an attempt at capturing the event, its participants, and the relations among them in a way that resembles MPEG-7-based semantic annotation. This kind of annotation deals
CHAPTER 4. CONCLUSION

primarily with the narrative world of the image, rather than with the spatial configuration of the figures. Incorporating this information in the annotation model, however, would be an interesting direction for future work.
Bibliography


