The Semantics of Image Annotation

Julia Bosque-Gil
Universidad Politécnica de Madrid
Brandeis University
jbosque@delicias.dia.fi.upm.es

James Pustejovsky
Computer Science Department
Brandeis University
jamesp@cs.brandeis.edu

Abstract

This paper presents a language for the semantic annotation of images, focusing on event types, their participants, and their spatial and orientational configurations. This language, ImageML, is a self-contained *layered specification* language, building on top of ISOspace, as well as some elements from Spatial Role Labeling and SpatialML. An annotation language characterizing such features surrounding an event and its various aspects could play a significant role in structured image retrieval, and a mapping of annotated semantic entities and the image's low-level features will likely assist event recognition and description generation tasks.

1 Introduction

The role of image annotation is becoming increasingly important in the context of algorithms that allow for efficient access and retrieval of images from large datasets; for this reason, it has become an active topic of research in both the computer vision and natural language processing communities. Keyword annotation (tagging) approaches include interactive annotation games (Von Ahn and Dabbish, 2004; Von Ahn et al., 2006; Ho et al., 2009) and automatic keyword annotation, where, given an image, the system provides the appropiate (or potential) labels that describe its content (Li and Fei-Fei, 2007; Luo et al., 2009; Feng and Lapata, 2010). On the other hand, efforts in the task of image caption generation have experienced a growth due to the advances in object recognition. Here, objects as well as relations among them have to be identified, and the output must be a grammatical (and, if possible, natural) sentence that correctly describes the image content (Kiros et al., 2014). Approaches include those of Farhadi et al. (2010); Elliott and Keller (2013); Kiros et al. (2014) and Karpathy and Fei-Fei (2014), among many others.

The current MPEG-7 format encodes several dimensions of information about image structure (visual features, spatio-temporal structure, decomposition in regions or shots, etc.) and semantic content by means of its descriptors (Martinez, 2004). Semantic annotation with MPEG-7 captures events represented in the image as well as participants (objects and agents), the time, location, etc., and annotation and retrieval tools based on this format were presented in Lux et al. (2003); Lux and Granitzer (2005) and Lux (2009). The use of ontologies and thesaurus in the annotation of the semantic content of an image has been developed in the art history domain in Hollink et al. (2003); Hollink (2006) and Klavans et al. (2008), as well as in the context of multimedia semantic indexing (Nemrava et al. (2008)).

This paper approaches the annotation of image content outside the task of automatic image caption generation. Even though MPEG-7 approaches capture information about the event, its participants and the relations among them, this annotation could be enriched to include aspects that go beyond the basic categories addressed so far (location, time, event, participants), such as: the spatial relations between participants, the motion of objects, the semantic role of participants, their orientation and frame of reference, the relations among events in the image, or the characterization of the image as a whole as prototypical, given the event in question. These aspects can be included following text annotation schemes such as SpatialML (Mani et al., 2010), ISOspace (Pustejovsky et al., 2011) and Spatial Role Labeling (Kordjamshidi et al., 2010). Pustejovsky and Yocum (2014) in fact adapt ISOspace to the annotation of the spatial configuration of objects in image captions, in particular to distinguish the way captions refer to the structure versus the content of the image. In this paper, we introduce ImageML for this purpose, and we describe how this richer information concerning the image can be incorporated as a self-contained *layered annotation*[1], making explicit reference to several embedded specifications, i.e., ISO-TimeML and ISOspace (ISO/TC 37/SC 4/WG 2 (2014); Pustejovsky et al. (2010)).[2]

2 Problems Posed by Images

Text-based image search assumes that images have an annotation of some kind or at least a text in which to perform the query, and, that the text of the web page on which the image appears is related to the image content. These two assumptions, however, do not always hold. Content-based image retrieval approaches the problem by recording the image's low-level features (texture, color layout, etc.) and semantic annotation of images aims to bridge the gap between those low-level features and the image semantic content.

However, efforts in keyword annotation, MPEG-7-based semantic annotation, and ontology-based annotation do not capture some aspects to which users might turn their attention when searching for an image. Although unstructured labels might be enough for image filtering or simple queries (*dog running in park*), more complex ones require a richer annotation that includes a description about the orientation of figures with respect to the viewer, the spatial relations among objects, their motion, appearance, or the structure of the event (including it sub-events) in which they might be involved; e.g., a user needs a picture of someone running towards the camera while listening to music.

MPEG-7-based annotation effectively captures the 'narrative world' of the image (Benitez et al., 2002), but does not provide a thorough annotation of the representation of figures or a characterization of their motion according to different frames of reference. Furthermore, image captions alone have a fixed frame of reference (viewer) and descriptions might refer both to image structure or image content; cf. (Pustejovsky and Yocum, 2014), which makes the annotation of this distinction an important task towards a more accurate image retrieval.

By capturing information about: (1) the event (type of event, any sub-events, any motion triggered by it, or any other event the image might refer to, if it is ambiguous); (2) the participants of the event (their type of entity, their semantic roles, their appearance, and their representation); and (3) the setting and the time of the depicted situation, ImageML would not only contribute to a more precise image querying capability, but it could also assist in event recognition and automatic caption generation tasks.

3 Annotating Spatial Relations in Images with ISOspace

The annotation of spatial information in text involves at least the following: a PLACE tag (for locations, entities participating in spatial relations, and paths); LINK tags (for topological relations, direction and orientation, time and space measurements, and frames of reference); and a SIGNAL tag (for spatial prepositions)[3]. ISOspace has been designed to capture both spatial and spatiotemporal information as expressed in natural language texts (Pustejovsky et al. (2012)). We have followed a strict methodology of specification development, as adopted by ISO TC37/SC4 and outlined in Bunt (2010) and Ide and Romary (2004), and as implemented with the development of ISO-TimeML Pustejovsky et al. (2005) and others in the family of SemAF standards.

There are four spatial relation tags in ISOspace, that are relevant to the definition of ImageML, defined as follows:

(1) a. QSLINK – qualitative spatial relations;
b. OLINK – orientation relations;
c. MLINK – dimensions of a region or the distance between them.

[1] Roser and Pustejovsky (2008); Lee (2013).

[2] The initial specification of a semantic annotation for images is first outlined in Bosque-Gil (2014).

[3] For more information, cf. Pustejovsky et al. (2012).

d. MOVELINK – for movement relations;

QSLINKs are used in ISOspace to capture topological relationships between tag elements captured in the annotation. The relType attribute values come from an extension to the RCC8 set of relations that was first used by SpatialML. The possible RCC8+ values include the RCC8 values Randell et al. (1992), in addition to IN, a disjunction of TPP and NTPP.

Orientation links describe non-topological relationships. A SPATIAL_SIGNAL with a DIRECTIONAL semantic_type triggers such a link. In contrast to qualitative spatial relations, OLINK relations are built around a specific frame of reference type and a reference point. The referencePt value depends on the frame_type of the link. The ABSOLUTE frame type stipulates that the referencePt is a cardinal direction. For INTRINSIC OLINKs, the referencePt is the same identifier that is given in the ground attribute. For RELATIVE OLINKs, the identifier for the viewer should be provided as to the referencePt. When the document type is IMAGE, all olinks are interpreted as relative FR relations (unless otherwise stated), with the "VIEWER" as the referencePt.

ISOspace also allows one to identify the source and type of the text being annotated. This is done with the document creation location (DCL) attribute. This is a distinguished location that serves as the "narrative or reference location". While useful for narratives and news articles, captions associated with images pose a different problem, in that the document describes a *representational artifact*,[4] such as an image or a Google *Street View* scene; hence, the document type is distinguished as an IMAGE. To account for this, (Pustejovsky and Yocum, 2014) introduce a new attribute, domain, which can take one of two values: STRUCTURE and CONTENT. This allows the spatial relations to differentiate the kinds of regions being identified in the caption. Furthermore, this means that the DCL can take two values: an *Image Structure Location*, for reference to the image as an object; and an *Image Content Location*, which is what the picture refers to (as in the default DCL for most texts).

4 The ImageML Model

In this section, we describe ImageML, a model for the semantic annotation of images. The conceptual schema provides a introduction to the information covered, its elements, and the relations among them.

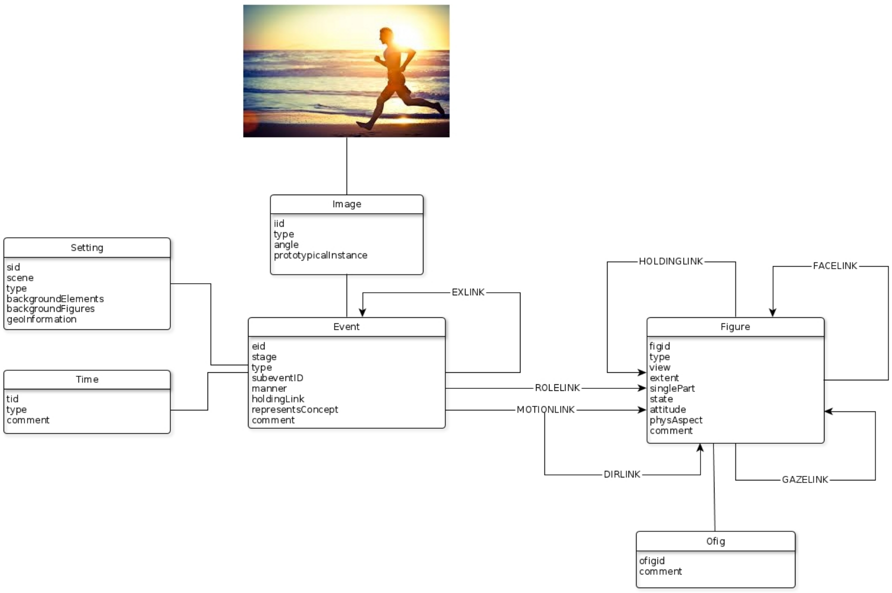

Figure 1: Conceptual Schema

[4] These are represented as $phys\_obj \bullet info$ complex types (dot objects), and inherit the properties of both type elements Pustejovsky (1995).

This annotation model is an attempt at capturing the semantic of images representing events, in contrast to images of landscapes and other non-dynamic representations. For this reason every annotated image includes at least one element of type EVENT.

The tags EVENT, FIGURE, OFIGURE, SETTING and TIME aim at encoding most of the information about the represented situation, both from a semantic perspective as well as from a formal one dealing with the specific way the elements are portrayed. In our view, participants of events have certain characteristics, such as their physical appearance or their type (an object, a person, etc.), are involved in events that affect their posture (eg. sitting, standing), might have a gesture that viewers interpret as them having an emotional attitude (which is valuable information for image descriptions), and are represented in a limited number of ways with respect to the viewer (back view, full body, etc.). Relation tags serve three purposes: first, capturing how a figure relates to an event and how the figure's specific representation is coupled with the characteristics of the motion involved in the event (if any); second, accounting for event ambiguity; and third, recording frequent sub-events of a main event which involve two participants (eg. holding, gazing, facing). We included the latter relations because they provide information that complements topological and spatial annotations without overcomplicating the annotation task. The modeling of these latter events as relations responds to the need to capture them in numerous images.

4.1 IMAGE

This tag records the type of image (e.g. photo) and the camera angle. Going back to Bloehdorn et al. (2005)'s knowledge base of prototypical instances, its attribute `prototypical` encodes whether an image could be considered a canonical instance of the event it depicts, which is valuable information for the event recognition task.

4.2 EVENT

The EVENT tag comes from TimeML (Pustejovsky et al., 2003; ISO/TC 37/SC 4/WG 2, 2012) and here it captures the activity, event, or change of state that is represented in the image. Its attribute `stage` encodes the phase of the event and the attribute `type` indicates whether the event is a sub-event of a main event or the main event itself, in which case its sub-events are also listed as values for the attribute `subevents`. Holding events that on first sight could have been thought of as EVENTs of type *sub-event* are here captured by links (HOLDINGLINK) to facilitate the annotation process. In this way, in a picture of someone taking notes holding a notebook and a pencil, only the event *take notes* would be recorded as EVENT, in this case of type *main* and with two HOLDINGLINKs, one for the pen and one for the notebook.

4.3 FIGURE and OFIGURE

FIGUREs are those objects in an image that are participants of an event or are involved in a holding relation. An object takes part in an event if it plays a semantic role (agent, theme, experiencer, etc.) in it, which is captured by the ROLELINK relation. This point is worth mentioning in order to distinguish FIGUREs from other objects that appear on the image but do not take part in any event, hence their description is outside the scope of this specification.

The type of object is encoded in the `type` attribute, which takes its values from the ACE Entity types[5], from MPEG-7 semantic descriptor values (*object* and *person*), and from SpatialML (*place*). The way the figure is portrayed with respect to the viewer in terms of a vertical axis (*front*, *profile-lateral*, etc.) and its perceivable extent (*whole*, *waist_up*, etc.) are recorded by the attributes `view` and `extent` respectively. Other properties such as physical appearance, attitude, or their state (e.g., open, broken, etc.) are also accounted for.

Figures not present in the picture but inferred by the reader when interpreting the image content are captured by the OFIGURE tag. Common examples are images in which a figure interacts with the

[5] ACE: Automatic Content Extraction 2008 Evaluation Plan (ACE08), `http://www.itl.nist.gov/iad/mig/ /tests/ace/2008/doc/`.

viewer (waving the hand at the camera, for instance), or close-up shots where the agent is not visible. An example of this is given below in Figure (2).

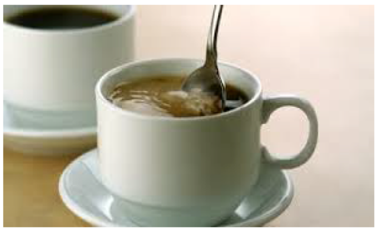

Figure 2: The OFIG is the agent of the event stir, the spoon is just the instrument.

4.4 SETTING and TIME

The SETTING tag aims at capturing information about the location of the events, any background elements or any background figures taking part in an event. Its attribute `figureID` distinguishes the overall setting of the events (e.g. a street) from specific objects in the image in which the event takes place (reading on *a bench* on the street), which are FIGURES with a role. The `scene` attribute records general aspects about the background of the image (e.g., *outdoors* and the attribute `type` encodes more specific information (e.g., *street*). Similarly, the TIME tag, inspired by TimeML TIMEX tag, encodes the time of the events and information deducible from the background.

4.5 ROLELINK and HOLDINGLINK

Kordjamshidi et al. (2010) introduce an annotation scheme similar in design to the task of semantic role labeling and classifies the arguments of spatial expressions in terms of their spatial roles. In this spirit, the tag ROLELINK addresses some spatial relations by indicating the source or goal of a movement, but it manly encodes the semantic roles participants of events play, turning to the semantic roles used in VerbNet (Schuler, 2005): *agent*, *recipient*, *instrument*, *experiencer*, etc. The HOLDINGLINK relation was introduced in section 4.2 and stands for holding sub-events: it links a figure (agent) that holds a figure (theme). Just as events, these relations have a `manner` attribute.

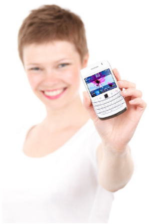

Figure 3: HOLDINGLINK expresses a sub-event hold, in which one figure holds another figure.

4.6 MOTIONLINK and DIRLINK

The tag MOTIONLINK is taken directly from ISOspace's MOVELINK tag. It associates to the event that triggers the motion, general information about the causer of the motion, the source and goal of it, and the path and medium through which the motion occurs. The orientation of the motion according to the different frames of reference is captured by the DIRLINK (direction link) relation, which combines attributes from SpatialML RLINKs and ISOspace OLINKs. The idea is to record fine-grained information about the orientation of the movement from the perspective of the object in motion, the causer of the movement and the viewer.

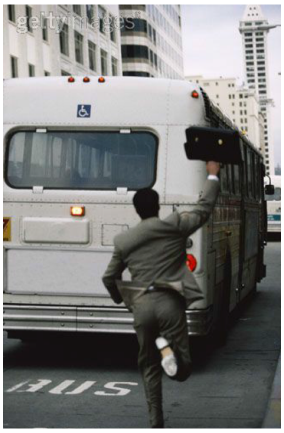

Figure 4: The relations MOTIONLINK and DIRLINK encode motion and direction of the event.

4.7 FACELINK AND GAZELINK

The relations FACELINK and GAZELINK draw upon the idea that eye gaze is an important semantic cue for understanding an image (Zitnick and Parikh, 2013). Facing and gazing could be thought of sub-events, but are here captured as links in a way resembling the relation HOLDINGLINK introduced earlier. Further, since a figure facing another figure does not imply that it is actually directing its gaze towards it, the FACELINK tag accounts for the way two figures are oriented towards one another, whereas the GAZINGLINK tag encodes eye-gaze relations between the two figures.

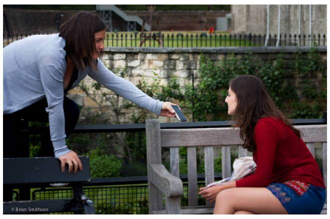

Figure 5: FACELINK captures facing relations between figures. A figure facing another may not be looking at it. For this reason, eye-gaze is encoded with GAZELINK.

4.8 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. The idea behind including both potential events is to allow for an association of the same low level features to both types of events in the context of automatic event recognition as well as for a retrieval of the image if the user searches for images of any of the two events.

5 Annotation Examples

To illustrate the descriptive nature of ImageML, let us consider an image that exploits many of the specification elements described above. This image is an instance of someone taking notes in a notebook.[6] The associated annotation identifies the event as "note-taking", along with the attributes of "holding a pen", the setting as being an interior location, the background being a bookshelf, and so on.

[6] Extracted from Google Image Search. Source: Flicker user Marco Arment (*marcoarment*).

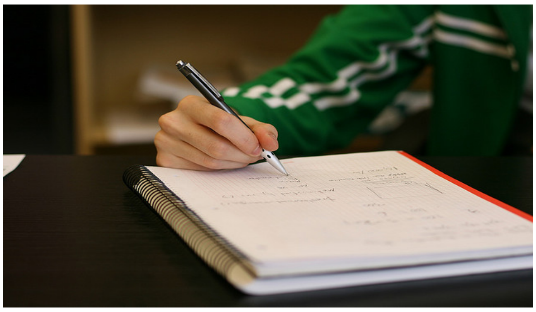

Figure 6: Brainstorming.

```
<IMAGE id="i0"  type="PHOTO" angle="NEUTRAL" prototypicalInstance="yes"/>
<FIGURE id="fig0"  type="PERSON" view="FRONT" extent="SINGLE_PART"
   singlePart="right hand and arm " state="" attitude="" physAspect="in a
   green sweatshirt" comment="">a student</FIGURE>
<FIGURE id="fig1"  type="OBJECT" view="PROFILE_LATERAL" extent="WHOLE"
   singlePart="" state="" attitude="" physAspect="black and silver"
   comment="">a pen</FIGURE>
<FIGURE id="fig2"  type="OBJECT" view="3/4" extent="INSIDE" singlePart=""
   state="open" attitude="" physAspect="spiral, square ruled" comment="">a
   college notebook</FIGURE>
<EVENT id="e0"  stage="DURING" type="MAIN_EVENT" subevent="e1" manner=""
   holdingLink="hl0" representsConcept="" comment="">take notes </EVENT>
<EVENT id="e1"  stage="DURING" type="SUB-EVENT" subevent="" manner=""
   holdingLink="" representsConcept="" comment=""> sit</EVENT>
<HOLDINGLINK id="hl0"  holderFigureID="fig0" heldFigureID="fig1"
   manner="in his right hand"/>
<ROLELINK id="rl0" figureID="fig0" eventID="e1" role="AGENT"/>
<ROLELINK id="rl1" figureID="fig1" eventID="e1" role="INSTRUMENT"/>
<ROLELINK id="rl2" figureID="fig2" eventID="e1" role="PLACE"/>
<SETTING id="l0"  figureID="" scene="INDOORS" type="FACILITY"
   backgroundElements="bookshelves" backgroundFigures=""
   geoInformation="">studying room</SETTING>
<TIME id="t0"  type="OTHER"></TIME>
```

Rather than merely annotating all events in the image equally, it is important to note that there is a topic event (“note-taking”), and that other salient eventualities, such the pen being held in a hand, etc., are captured as relational attributes to the main event. This would not be sufficient for a general event description protocol, such as that promoted in ISO-TimeML, but for image descriptions, it appears particularly well-suited, at least in the context of images that we have so far studied. Obviously, this is an issue that deserves further empirical study.

6 Conclusion

In this paper we have presented ImageML, a model for the semantic annotation of images which draws largely upon ISOspace, as well as some aspects of Spatial Role Labeling and SpatialML, to capture fine-grained information about the events depicted in an image, the motion involved (described from different frames of reference), as well as information about the participants, their orientation, and the relations among them. The setting and time of the situation are also accounted for. By its very design, ImageML is a layered annotation, incorporating elements and values from the embedded specification languages of ISOspace and ISO-TimeML.[7]

While not yet created, a database of images annotated with this information along with the spatial configuration of objects should be of potential use to structured image retrieval, event detection and recognition, and automatic image caption generation. We are currently pursuing the creation of such a corpus.

Acknowledgements

Julia Bosque-Gil was partially supported by the LIDER project: “Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe”, a FP7 project reference number 610782 in the topic ICT-2013.4.1: Content analytics and language technologies. James Pustejovsky was partially supported by grants from NSF’s IIS-1017765 and NGA’s NURI HM1582-08-1-0018. All errors and mistakes are, of course, the responsibilities of the authors.

References

Benitez, Ana B, Hawley Rising, Corinne Jorgensen, Riccardo Leonardi, Alessandro Bugatti, Koiti Hasida, Rajiv Mehrotra, A Murat Tekalp, Ahmet Ekin, and Toby Walker (2002). Semantics of multimedia in mpeg-7. In *Proceedings of the International Conference on Image Processing*, Volume 1, pp. I–137. IEEE.

Bloehdorn, Stephan, Kosmas Petridis, Carsten Saathoff, Nikos Simou, Vassilis Tzouvaras, Yannis Avrithis, Siegfried Handschuh, Yiannis Kompatsiaris, Steffen Staab, and Michael G Strintzis (2005). Semantic annotation of images and videos for multimedia analysis. In *The semantic web: research and applications*, pp. 592–607. Springer.

Bosque-Gil, Julia (2014). *A Model for the Semantic Annotation of Images*. MA Thesis, Brandeis University.

Bunt, H. (2010). A methodology for designing semantic annotation languages exploiting syntactic-semantic iso-morphisms. In *Proceedings of ICGL 2010, Second International Conference on Global Interoperability for Language Resources*.

Elliott, Desmond and Frank Keller (2013). Image Description using Visual Dependency Representations. In *EMNLP*, pp. 1292–1302.

Farhadi, Ali, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth (2010). Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pp. 15–29. Springer.

Feng, Yansong and Mirella Lapata (2010). Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 831–839. Association for Computational Linguistics.

Ho, Chien-Ju, Tao-Hsuan Chang, Jong-Chuan Lee, Jane Yung-Jen Hsu, and Kuan-Ta Chen (2009). KissKissBan: a competitive human computation game for image annotation. In *Proceedings of the ACM SIGKDD workshop on human computation*, pp. 11–14. ACM.

Hollink, Laura (2006). *Semantic annotation for retrieval of visual resources*. Ph. D. thesis.

Hollink, Laura, Guus Schreiber, Jan Wielemaker, Bob Wielinga, et al. (2003). Semantic annotation of image collections. In *Knowledge capture*, pp. 41–48.

[7] Features from SpatialML are already incorporated into ISOspace, and the relations in Spatial Role Labeling are captured through the relation tags in ISOspace.

Ide, N. and L. Romary (2004). International standard for a linguistic annotation framework. *Natural Language Engineering 10*(3-4), 211–225.

ISO/TC 37/SC 4/WG 2, Project leaders: James Pustejovsky, Kiyong Lee (2012). Iso 24617-1:2012 language resource management - part 1: Time and events (iso-timeml). ISO/TC 37/SC 4/WG 2.

ISO/TC 37/SC 4/WG 2, Project leaders: James Pustejovsky, Kiyong Lee (2014). Iso 24617-7:2014 language resource management - part 7: Spatial information (isospace). ISO/TC 37/SC 4/WG 2.

Karpathy, Andrej and Li Fei-Fei (2014). Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*.

Kiros, Ryan, Ruslan Salakhutdinov, and Richard S Zemel (2014). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Klavans, J., C. Sheffield, E. Abels, J. Beaudoin, L. Jenemann, J. Lin, T. Lippincott, R. Passonneau, T. Sidhu, D. Soergel, et al. (2008). Computational Linguistics for Metadata Building: Aggregating Text Processing Technologies for Enhanced Image Access. In *OntoImage 2008: 2nd Workshop on Language Resources for Content-Based Image Retrieval, LREC*, pp. 42–46.

Kordjamshidi, Parisa, Marie-Francine Moens, and Martijn van Otterlo (2010). Spatial role labeling: Task definition and annotation scheme. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 413–420.

Lee, Kiyong (2013). *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, Chapter Multi-layered Annotation of Non-textual Data for Spatial Information, pp. 15–24. Association for Computational Linguistics.

Li, Li-Jia and Li Fei-Fei (2007). What, where and who? Classifying events by scene and object recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8. IEEE.

Luo, Jie, Barbara Caputo, and Vittorio Ferrari (2009). Whos doing what: Joint modeling of names and verbs for simultaneous face and pose annotation. In *Advances in Neural Information Processing Systems*, pp. 1168–1176.

Lux, Mathias (2009). Caliph & Emir: MPEG-7 photo annotation and retrieval. In *Proceedings of the 17th ACM international conference on Multimedia*, pp. 925–926. ACM.

Lux, Mathias, Jutta Becker, and Harald Krottmaier (2003). Semantic Annotation and Retrieval of Digital Photos. In *CAiSE Short Paper Proceedings*.

Lux, Mathias and Michael Granitzer (2005). Retrieval of MPEG-7 based semantic descriptions. In *BTW-Workshop "WebDB Meets IR"*, Volume 11.

Mani, Inderjeet, Christy Doran, Dave Harris, Janet Hitzeman, Rob Quimby, Justin Richer, Ben Wellner, Scott Mardis, and Seamus Clancy (2010). Spatialml: annotation scheme, resources, and evaluation. *Language Resources and Evaluation 44*, 263–280. 10.1007/s10579-010-9121-0.

Martinez, Jose Maria (2004). MPEG-7 overview (version 10), ISO. Technical report, IEC JTC1/SC29/WG11.

Nemrava, Jan, Paul Buitelaar, Vojtech Svatek, and Thierry Declerck (2008). Text mining support for semantic indexing and analysis of a/v streams. In *OntoImage 2008*. ELDA.

Pustejovsky, James (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Pustejovsky, James, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering 3*, 28–34.

Pustejovsky, James, Robert Knippen, Jessica Littman, and Roser Saurí (2005, May). Temporal and event information in natural language text. *Language Resources and Evaluation 39*, 123–164.

Pustejovsky, James, Kiyong Lee, Harry Bunt, and Laurent Romary (2010). Iso-timeml: A standard for annotating temporal information in language. In *Proceedings of LREC*, pp. 394–397.

Pustejovsky, James, Jessica Moszkowicz, and Marc Verhagen (2012). A linguistically grounded annotation language for spatial information. *TAL 53*(2).

Pustejovsky, James, Jessica L Moszkowicz, and Marc Verhagen (2011). ISO-Space: The annotation of spatial information in language. In *Proceedings of the Sixth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pp. 1–9.

Pustejovsky, James and Zachary Yocum (2014). Image Annotation with ISO-Space: Distinguishing Content from Structure. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Randell, David, Zhan Cui, and Anthony Cohn (1992). A spatial logic based on regions and connections. In M. Kaufmann (ed.), *Proceedings of the 3rd Internation Conference on Knowledge Representation and Reasoning*, San Mateo, pp. 165–176.

Roser, Saurí and J Pustejovsky (2008). From structure to interpretation: A double-layered annotation for event factuality. In *Proceedings of the 2nd Linguistic Annotation Workshop*.

Schuler, Karin Kipper (2005). Verbnet: A broad-coverage, comprehensive verb lexicon.

Von Ahn, Luis and Laura Dabbish (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 319–326. ACM.

Von Ahn, Luis, Ruoran Liu, and Manuel Blum (2006). Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pp. 55–64. ACM.

Zitnick, C Lawrence and Devi Parikh (2013). Bringing semantics into focus using visual abstraction. In *Conference on Computer Vision and Pattern Recognition (CVPR), 2013*, pp. 3009–3016. IEEE.