Spatio-Temporal Visual Ontology

Joanna Isabelle Olszewska
http://compeng.hud.ac.uk/scomjio

School of Computing and Engineering
University of Huddersfield, UK

In order to semantically describe and structure the visual content of images and video scenes captured by cameras with arbitrary resolution and unknown calibration properties, we propose a spatio-temporal visual ontology (STVO).

In Artificial Intelligence (AI), an ontology is an explicit specification of a conceptualization [4] and thus consists of a set of semantic concepts as well as the relations between them. In Computer Vision, the proposed conceptualizations of visual scenes could refer to low-level features [3], objects [6], spatial relations between objects [5] or are event-based specific domain ontologies [1]. In fact, previous works have been focused on the study of scenes recorded by static cameras and each of these approaches only partially covers the entire range of general semantic concepts necessary for the interpretation of dynamic scenes captured by mobile cameras.

Our STVO ontology has been conceived to be well suited for the general case of automatic understanding of visual scenes recorded by either stationary and moving cameras and provides an innovative granular conceptualization of the visual domain as well as a set of original inter- and intra-object spatio-temporal relations and visual properties defined for this framework [9].

STVO which is presented in Fig. 1 was designed by using a top-down approach [2], in the way its hierarchical concepts specifications map the granularity of the visual scene. The depth of STVO's ontological tree is five, corresponding to visual levels of video sequence, scene, object, parts of objects and attributes, respectively, and enables a higher computational efficiency in terms of navigability than deeper ontologies such as [6], while the structure of STVO allows clarity, coherence as well as an efficient scalability and usability.

To develop STVO domain knowledge, we have adopted the OWL standard language [8] to express the complex semantic entities and their relations. Its description logic (DL) allows us to perform automated reasoning on it and to formulate queries related to both semantic and numerical properties of visual scenes.

In our ontology, an object of interest is conceptualized in terms of color, shape, position and other attributes mapping the visual observation domain properties. In particular, the color concept is based on the 16 color keywords defined in [12] as well as on its related explicit (R,G,B) value in the red (R), green (G), blue (B) color space. In the experiments carried out on real-world scenes acquired by mobile camera [9], this choice gave satisfactory results and ensures computational effectiveness. However, these semantic concept values could be extended to more keywords if any application requires that.

In STVO, we introduce the concept of the color of an object (*Object_Color*) containing p parts P_i *with* $(i = 1, ..., p)$ as follows:

$$\begin{aligned} Object\_Color \equiv\ & Part\_Color.P_1 \\ & \sqcup ... \sqcup Part\_Color.P_p \end{aligned} \tag{1}$$

with

$$\begin{aligned} Part\_Color \equiv\ & Color \\ & \sqcap \exists hasRChannelValue_{=R_p} \\ & \sqcap \exists hasGChannelValue_{=G_p} \\ & \sqcap \exists hasBChannelValue_{=B_p} \end{aligned} \tag{2}$$

In this example, $=R_p$, $=G_p$, and $=B_p$ are each a predicate over the number domain $[0,255]$. The DL definition of the concept of the object color proposed in Eq. (1) takes into account the color of the different parts and not necessarily the average value over the whole detected object. Hence, this novel definition is more discriminant for objects with inhomogeneous color properties as it occurs in real-world scenes [9].

On the other hand, features such as texture are not relevant in STVO because of the blurring effect of the camera movement. Moreover, as we assume the general case of the unknown calibration of the mobile camera, properties such as object motion are not considered.

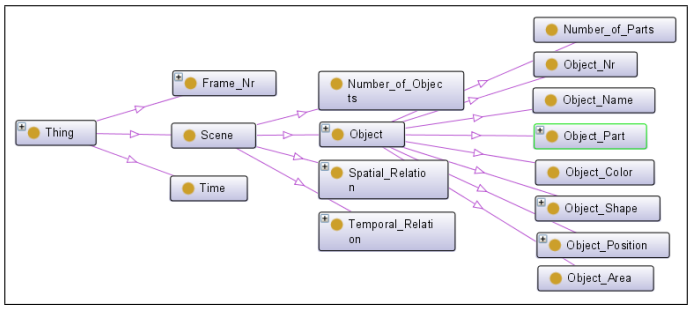

Figure 1: Spatio-Temporal Visual Ontology (STVO).

For the spatial relations between and within objects, we define in STVO three families: (i) the topological relations, (ii) the directional relative positions, and (iii) the relative distances, sizes, and areas.

For (i), we adopt the RCC-8 model [10] consisting of eight basic relations between two objects: disconnected (DC), externally connected (EC), equal (EQ), partially overlapping (PO), tangential proper part (TPP), tangential proper part inverse (TPPi), non-tangential proper part (NTPP), non-tangential proper part inverse (NTPPi).

For (ii), we introduce in our ontology original semantically meaningful concepts in regards to the *o'clock* notion [9] in order to provide a unified framework for inter- and intra-object spatial relations and to reduce the uncertainty on the relative positions between two objects. Our new formal specification of the relative directional relations is based on the clock space that numbers the scene plane as a clock's face. Indeed, in real-world crowded scenes, saying that an object is at the right of another could certainly be not enough discriminant as it could be many of them. Indicating the o'clock relative position could reduce the uncertainty related to the directional relative positions among objects.

As an example of how we have formalized the *o'clock* concept, we define the *2 o'clock* concept in DL as follows. Let O_{REF} be the object of reference, and let the object we want to know its relative directional relation with be called O_{REL}. Let *Angle* be the relative angle between the line O_{REF}–O_{REL} and the X axis of the image plane. Then:

$$\begin{aligned} isAt2clockOf \sqsubseteq\ & Spatial\_Relation \\ & \sqcap Relative\_Distance\_Relation \\ & \sqcap \exists O_{REF} \\ & \sqcap \exists O_{REL} \\ & \sqcap \exists Angle2clock \\ & \sqcap \exists inverse.isAt8clockOf \end{aligned} \tag{3}$$

with

$$\begin{aligned} Angle2clock \equiv\ & Angle \\ & \sqcap \exists angle.value_{\leq \frac{\pi}{6}} \\ & \sqcap \exists angle.value_{>0} \end{aligned} \tag{4}$$

Equation (4) denotes the set of angles which have values lower than or equal to $\pi/6$ and higher than 0. In this example, $_{\leq \frac{\pi}{6}}$ is a predicate over the real number domain $\mathbb{R}$. The *2 o'clock* concept has the inverse property *8 o'clock*. Indeed, if a relative object (O_{REL}) is at *2 o'clock* of a reference object (O_{REF}), the reference object is at *8 o'clock* of the relative object (O_{REL}). Moreover, the traditional relative intra-object relations (*Above_Of*, *Below_Of*, *Left_Of*, *Right_Of*) are embedded in our concept. Indeed, e.g. the property *isAt3clockOf* could be seen as equivalent to the traditional directional *Right_Of* relation.

Semantically meaningful spatial relative relations between p parts P_i of an object are introduced for the first time in an ontology in our work [9] and we called them intra-object relations. We formalize them using our new *o'clock* concept which does not induce an arbitrary division of

the target but a partition taking automatically into account object's concavities and convexities.

For (iii), we define the concept of being *close* or *far* in terms of a proportion between r, the distance in the image plane between O_{REF} and O_{REL} and the size of the width and height of the scene image. We also propose the semantic concept of relative area and size of objects [9]. These notions could help for example in scene understanding in situations such as an object is on the first plane, or there is a close-up on a target object.

For the temporal relations, we focus on the main notions *hasAppeared*, *hasDisappeared*, and *isInScene*. These first two relations are related to the *start* and *end* ones [7], [11]. The DL description of our STVO property *isInScene* is as follows:

$$\begin{aligned} isInScene \sqsubseteq\ & Temporal\_Relation \\ & \sqcap hasAppeared \\ & \sqcap \neg hasDisappeared \end{aligned} \tag{5}$$

where the object properties are defined through the existence or not of the *Object_Position* in a scene.

Since many of STVO relations take the form of logic rules on which automatic reasoning can take place, this leads to the automatic understanding of complex dynamic scenes as demonstrated in [9].

STVO allows not only simple queries like in state-of-art ontologies based on only one type of relations such as [5], but also complex queries based on multiple criteria (spatial, temporal, shape, appearance) providing significant improvements by reducing the set of the possible answers and thus enabling more precise answers than those obtained e.g. through [6] which is lacking of temporal relations.

STVO provides also inter-object relations (*o'clock* ones) which outperform the state-of-art spatial relations (right, left, etc.) used e.g. in [5] because of a reduction of ambiguity in situations such as crowd. Moreover, based on this *o'clock* notion, STVO includes intra-object relations which are unique as none are specifically proposed in the literature and which lead to a better characterization of the color semantic definition (set of colors of all parts of the object) than the state-of-art definitions (based on a single average value [3]), specially in case of complex objects like a player with shorts, shirt, etc. of different colors.

In conclusion, our STVO ontology can semantically characterize each visual scene and its components e.g. objects of interest and support reasoning about photometric, geometrical and our semantically meaningful spatio-temporal relations between and within the multiple observed objects or parts of them for the effective interpretation of visual information.

[1] M. Bertini, A. Del Bimbo, G. Serra, C. Torniai, R. Cucchiara, C. Grana, and R. Vezzani. Dynamic pictorially enriched ontologies for digital video libraries. *IEEE Multimedia*, 16(2):42–51, April 2009.

[2] O. Corcho, M. Fernandez-Lopez, and A. Gomez-Perez. Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data and Knowledge Engineering*, 46(1):41–64, July 2003.

[3] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, and M. G. Strintzis. Knowledge-assisted semantic video object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1210–1224, October 2005.

[4] T. R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5-6):907–928, November 1995.

[5] C. Hudelot, J. Atif, and I. Bloch. Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets and Systems*, 159(15):1929–1951, August 2008.

[6] N. Maillot and M. Thonnat. Ontology based complex object recognition. *Image and Vision Computing*, 26(1):102–113, July 2008.

[7] G. Nagypal and B. Motik. A fuzzy model for representing uncertain, subjective and vague temporal knowledge in ontologies. In *Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics*, pages 906–923, November 2003.

[8] B. Neumann and R. Moeller. On scene interpretation with description logics. *Image and Vision Computing*, 26(1):114–126, January 2008.

[9] J. I. Olszewska and T. L. McCluskey. Ontology-coupled active contours for dynamic video scene understanding. In *Proceedings of the IEEE International Conference on Intelligent Engineering Systems*, pages 369–374, June 2011.

[10] D. A. Randell, Z. Cui, and A. G. Cohn. A spatial logic based on regions and connection. In *Proceedings of the International Conference on Knowledge Representation and Reasoning*, pages 165–176, October 1992.

[11] J. G. Stell and M. West. A four-dimensionalist mereotopology. In *Proceedings of the International Conference on Formal Ontology in Information Systems*, pages 261–272, November 2004.

[12] SVG 1.1, W3C TR, Recognized color keyword names. Available online: `http://www.w3.org/TR/SVG/types.html#ColorKeywords`, June 2010.