The Baltic International Yearbook of
Cognition, Logic and Communication

November 2015 Volume 10: *Perspectives on Spatial Cognition*
pages 1-12 DOI: http://dx.doi.org/10.4148/1944-3676.1104

TAO WANG
University of Bremen

HUI SHI
University of Bremen

DESCRIBING IMAGES USING A MULTILAYER FRAMEWORK BASED ON QUALITATIVE SPATIAL MODELS

ABSTRACT: To date most research in image processing has been based on quantitative representations of image features using pixel values, however, humans often use abstract and semantic knowledge to describe and analyze images. To enhance cognitive adequacy and tractability, we here present a multilayer framework based on qualitative spatial models. The layout features of segmented images are defined by qualitative spatial models which we introduce, and represented as a set of qualitative spatial constraints. Assigned different semantic and context knowledge, the image segments and the qualitative spatial constraints are interpreted from different perspectives. Finally, the knowledge layer of the framework enables us to describe the image in a natural way by integrating the domain-specified semantic constraints and the spatial constraints.

1. MOTIVATION

As a basic carrier of visual information, images have been widely used and studied. Digital images are technically considered pixel matrices

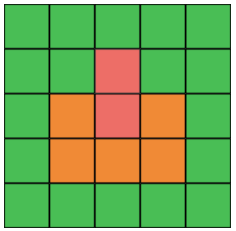
with different precisions and color spaces, and there is lots of work dedicated to extracting and describing quantitative image features using pixel values. Many global and local descriptors have been developed, such as LBP (Wang & He 1990), SIFT (Lowe 1999), HOG (Dalal & Triggs 2005), SURF (Bay et al. 2008). However, the understanding achieved as a result of human vision normally goes beyond what can be extracted from pixels insofar as human vision extracts different information or features from different perspectives and based on background considerations or knowledge.

Take the image below as an example. If people are asked to describe images like the scene shown in Fig. 1(a), possible answers could be, (1) a woman in red is riding a brown horse, (2) a person is on the back of a horse, or (3) there is something in red on something in brown, depending on their personal background knowledge and the situational context.

(a)

$$OIM(A,B) = \begin{bmatrix} 0 & 2 & 0 \\ 1 & 3 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

(b)



(c)

Figure 1: (a) a scene, (b) an OIM relation, and (c) a qualitative layout representation.

This example shows that humans often use semantic and background knowledge in image analysis, instead of relying only on quantitative data of the sort extracted by most of the current feature-extraction algorithms. Nowadays most image analysis research uses quantitative descriptions of image features based on pixel values. There is thus a need for the introduction of qualitative and semantic approaches to image representation in order to bridge the gap between the ways of processing images used by humans and those employed by existing algorithms. Zhang et al. (2012) is a good review of automatic image annotation techniques, a field where issues of image segmentation have drawn a great deal of attention, along with issues such as color features (e.g., color histogram), texture features (e.g., model-based), shape features (e.g.,contour/region-based) and spatial relationships (e.g., 2D string).

Some researches focus on describing images with high level semantic knowledge. Although background knowledge is extremely important in human image analysis, its formal modeling is still a big challenge. Techniques such as *single word vector spaces* are widely used, and many researchers are also trying to identify compositional meaning representations for longer phrases, or even to generate sentences. R. Socher et.al. introduced the dependency-tree recursive neural networks model in order to retrieve images described by sentences and to generate sentence descriptions for those images Socher et al. (2014). While this approach has yielded very encouraging results, it still produces some unsatisfactory results as well. As shown in Fig.1, some illogical results have been generated by the current fully learning-based method as a result of the fact that the domain knowledge and spatial restrictions are not carefully considered.

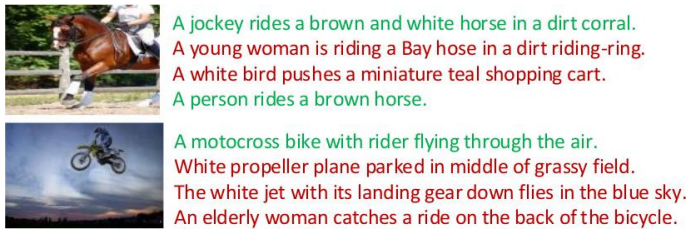


Figure 2: Failure examples from Socher et al. (2014), the generated sentences in red are rather illogical to humans.

On the other hand, some work explores comprehensive text-to-graphics systems as well. Based on Frame Semantics that represents conceptual and graphical relations, B.Coyne et. al. proposed Vignette Semantics, which is aimed at relating language to a grounded (e.g., graphical representation) semantics Coyne et al. (2011). Though VigNet takes internal structures and lexical spatial relations between objects in a scene into account, it does not make use of formally defined spatial models, which means that logical conflicts and faulty representations are still difficult to avoid. These failings further point out the necessity of introducing well-defined spatial models. There are several well studied qualitative spatial models (cf. Allen (1983); Goyal & Egenhofer (2000)), which provide plausible foundations for describing spatial relations of image segments, such that answers like the third one in the first example could become possible. Hence, we are concerned in this paper with the representation of segmented images using qualitative spatial models. Furthermore, while integrated with online learning techniques and domain specified training sets as in Socher et al. (2014), on our approach higher level descriptions are also going to be reliable.

2. A MULTILAYER FRAMEWORK FOR IMAGE DESCRIPTION

In this section we present a *multilayer framework for image description* intended to overcome the drawbacks that result when the domain knowledge and spatial restrictions are not taken into account. Generally, the framework introduces qualitative spatial models to formalize the spatial layout of the images, and creates a uniform architecture for computational and cognitive systems by adapting the basic ideas of *concepts as heterogeneous proxytypes* proposed by Lieto (2014).

The framework consists of three layers. The first layer is the visual layer at which the image is decomposed into segments, and the spatial relations between each of them are abstracted according to the qualitative spatial models that we introduce. Also at this level, visual and property features will be extracted depending on their importance in the later processing steps. In the second layer, named the concept layer, the segments are recognized and/or classified by the pre-trained system based on the visual features that have been extracted at the first layer, this results in the corresponding concept annotations being assigned to the segments. Meanwhile, the formalized spatial constraints are interpreted based on the primitive or reasoning-based semantic interpretations defined by the adopted spatial models. The domain specified description would be generated in the knowledge layer, i.e. the third layer, by taking both the domain specified (or situational) semantic constraints and the spatial constraints into consideration. Fig.3 gives an overview of the framework.

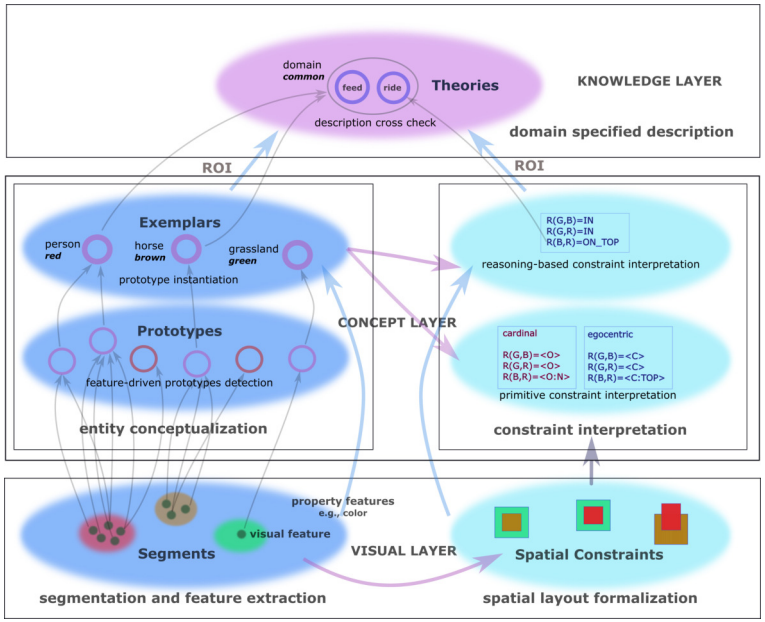


Figure 3: A multilayer framework for image description.

2.1. *Layer I. the Visual Layer*

Though our physical world is three-dimensional, setting aside time for the moment, images are basically a two-dimensional projection profile of it. Consequently, it is impractical to model a single image with accurate 3D perspective. In this paper, we treat images as the orthographic projection of the visual physical scene only, ignoring the 3D transformations resulting from the subject's point of view.

Images can be decomposed into sets of regions. Much current research is working on these issues, including segmentation, salient region extraction, and so on. Qualitative models can be used to bridge the semantic gap between the low-level features of images and human cognition. Well-defined qualitative spatial models have been proposed and investigated. These models are normally more cognitivly adaptive,

and could develop or already have their respective semantic primitive and/or reasoning-based interpretations, forming different and interdisciplinary research communities. Some also include consistency checking features and reasoning algorithms that can be used for later extension or model transformation.

Image segmentation techniques usually decompose an image into a set of non-empty, bounded and connected segments, i.e. simple regions, so qualitative spatial models for two dimensional regions could be adopted to represent their spatial layouts. We employ the objects interaction matrix model OIM proposed by Schneider et.al. Schneider et al. (2012), a further development of the direction relation matrix model Goyal & Egenhofer (2000) for representing direction relations in spatial regions, for example, in geographic information systems.

Suppose $S = \{s_i\}_{i=1}^{k}$ is the set of the segments of image I. Taking s_i and s_j from S, their minimal bounding boxes divide the space into a $n \times m$ $(n, m \in \{1, 2, 3\})$ matrix, depending on their relative positions (see Schneider et al. (2012)). The interaction relation of s_i with respect to s_j is defined as a $n \times m$ $(n, m \in \{1, 2, 3\})$ matrix, denoted as $\mathsf{OIM}(s_i, s_j) = M_{ij}$, where a field in M_{ij} is 1(2) if only $s_i(s_j)$ occurs in the corresponding field of the space matrix, 3 if the space matrix field is occupied by both s_i and s_j, and otherwise it is 0. For instance, the relation of the bounding boxes of the horse with respect to the rider (woman) is given in Fig. 1(b).

The *layout of segment* s_i in image I is then the set of region interaction relations of s_i with respect to all other segments, i.e., $\mathscr{L}(s_i) = \{M_{ij} | j \neq i, s_j \in S\}$. Moreover, the *layout of image* I is the union of the layouts of its segments, written as $R = \bigcup_{i=1}^{k} \mathscr{L}(s_i)$. Thus, according to the spatial layout formalization process, the given segmented image could be abstracted into a pair $I =< S, R >$. Fig.4 shows an example.

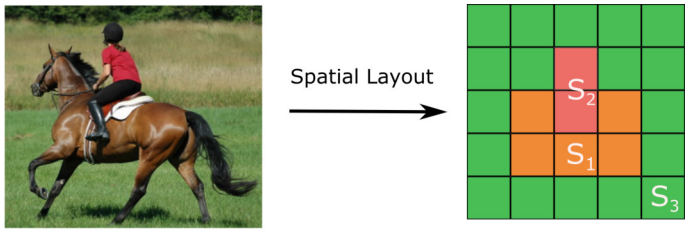


Figure 4: An example of the visual layer.

2.2. *Layer II. the Concept Layer*

The aforementioned visual layer establishes an organized picture of the input image, i.e. $I =< S,R >$, that provides the formal data for the concept layer that follows. There are two focuses at the concept layer: entity conceptualization of the segments and constraint interpretation on the basis of the qualitative spatial models.

According to the *Heterogeneous Hypothesis of Concepts* proposed by A. Lieto in Lieto (2014), the proxytype theories of concepts create an uniform interface which could be shared by the three different types of concepts that are assumed to exists: prototypical concepts, exemplar based concepts, and theory-theory concepts. In our framework, we basically adopt this taxonomy about the heterogeneous concepts. Referring to the same concept entity, there can be different types of concepts constituting different bodies of knowledge, and it can be represented by applying different ontologies and models that are already available or to be developed. In particular, regarding the visual layer of our framework, the entity conceptualization process would begin with feature-driven prototype-detection, and generate corresponding (semantic) exemplars by taking instantiated prototypes (e.g., recognized labels) and additional property features (e.g., color) both into consideration. A typical prototype network is built by supervised training algorithms based on the visual features extracted from a manually labeled set, while the exemplars usually depend on the assigned labels and employed ontologies. For example, for the segment S_2 shown in Fig.4, suppose there are two kinds of ontologies, *sports* and *common*, according to which the segment S_2 could be labeled with *'rider'* or *'person'*, respectively; meanwhile, referring to different locations in the taxonomy provided by an ontology, S_1 could be labeled with *'animal'* or *'horse'*. Moreover, the combination of the ontological and attributive interpretations could furnish more particular descriptions, like *'brown horse'*.

Constraint interpretation is twofold, involving both the primitive interpretation and the reasoning-based interpretation. Usually the spatial models have their interpretations, which can be used for description directly, while reasoning based interpretation could be introduced as well. The representation of image layouts using the qualitative spatial model OIM makes several qualitative analyses of segmented images possible. The most relevant one is to describe the relative spatial locations between image segments in a natural way, using their qualitative layout representations. Further, a *qualitative image reconstruction* algorithm, based on the theoretical work of Liu et al. (2010) and Li (2013), has been developed, which generates a grid model from the qualitative layout definition of an image in OIM, and the reconstructed spatial layouts enable interpretations involving more than two segments. Fig. 1(c) shows the qualitative layout representation of the scene in Fig. 1(a) generated by the reconstruction algorithm. With different constraints based on interpretation and perspective, the qualitative layouts can be naturally described from different points of view, including cardinal directions (e.g., *north* and *northwest*), egocentric directions (e.g., *left* and *in front*), and connections (e.g., *overlap* and *away from*). Considering the layout representation in Fig. 1(c), possible descriptions are, for example, "segment S_1 (*horse*) is *at the center of* segment S_3 (*grassland*)"; "segment S_2 (*woman/rider*) is *at the center and the top of* segment S_3 (*horse*) ". If the reasoning-based interpretations are considered, we then have "segment S_2 (*woman/rider*) is *on* segment S_1 (horse)". Furthermore, a spatial layout comparison test with 140 man-

ually segmented images from the IAPR TC-12 dataset Escalante et al. (2010) has been carried out. The experiment results show that the qualitative layouts reconstructed by the algorithm are consistent with those that have been manually created.

2.3. Layer III. the Knowledge Layer

The knowledge layer mainly aims to contribute the high level interpretations that are clear and natural for human understanding. Toward this goal, a domain knowledge network based on existing high level semantic techniques would be established, with which the framework could supervise the selection of an ontology or decide the suitable level of generality within an ontology.

On one hand, a common way to build such a domain knowledge network is to train an artificial neural network on a manually collected and labeled dataset which consists of a mass of images with their corresponding descriptions, such as keywords, phrases, or even sentences. These kinds of learning-based methods have been studied deeply, nevertheless, as mentioned in the motivation section, knowledge networks generated in this way lack semantic logic and strict domain limits in some cases. Thus, the learned knowledge network also needs to be integrated with domains and ontologies in order to restrict the learned-knowledge network to using annotations from an existing ontology. On the other hand, activity modeling based on qualitative spatial models and the situational context data (e.g., domains) have been explored, and the modeled activities are connected to the corresponding nodes of the knowledge network, and could be used to filter and verify the domain specified descriptions.

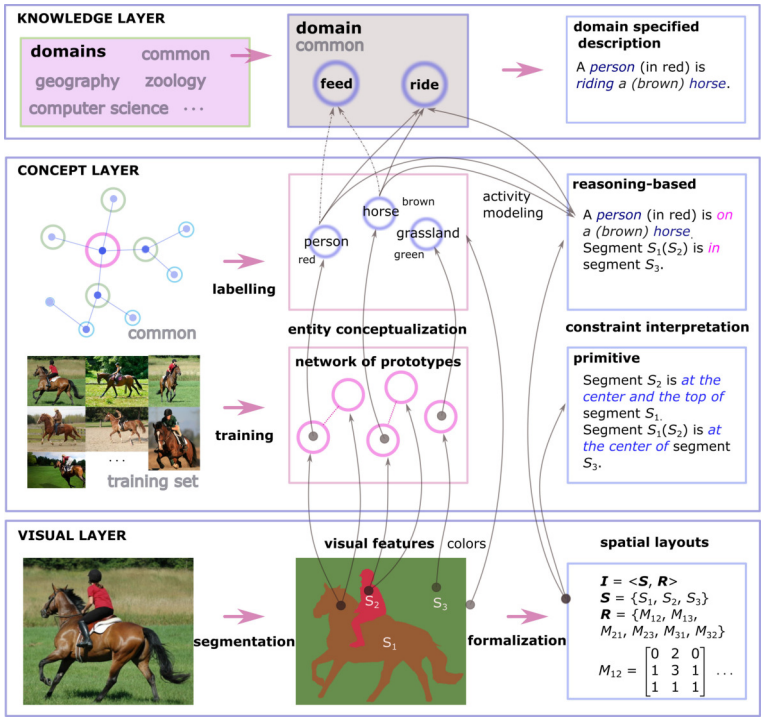


Figure 5: An example of the multilayer framework for image description.

3. DISCUSSION

Fig. 5 shows an example of the working flow of the multilayer framework for image description. The framework basically shares the common methods with traditional quantitative algorithms in image segmentation, feature extraction and entity conceptualization procedures. However, the framework also abstracts the spatial layout of the image segments formally according the qualitative spatial models that have been introduced; in this work, we adopt the objects interactions model. The qualitative spatial models not only provide the formal basis for fea-

tures such as consistency checking to ensure the validity of the spatial layouts, but also offer the the necessary primitive interpretations and could be used to construct reasoning-based interpretations and to model activities. By taking both domain restrictions and spatial constraints into account, the knowledge layer could describe the given image in a more reasonable manner. Taking the input image of Fig.5 as an example, the traditional quantitative algorithms might obtain two descriptions, for example "a person is feeding a horse" and "a person is riding a horse"; while our framework would filter out the first one because of its unmatched spatial layout.

4. CONCLUSION

In this paper we proposed a multilayer framework for image description aiming to bridge the gap between the ways of processing images used by humans and those found in existing quantitative algorithms. Though the traditional quantitative image description algorithms achieve very encouraging performances, they also generate illogical results because the domain knowledge and spatial restrictions are not carefully considered. The framework proposed here introduces qualitative spatial models to formalize the spatial layout of the images, thus recording the spatial restrictions, and adopts the taxonomy of heterogeneous hypothesis of concepts, thus creating a uniform interface which could be shared by different types of concepts to keep the domain restrictions and semantic logic structures. Finally, the framework establishes a domain knowledge network integrating the domain-specified semantic constraints and the spatial constraints, in order to describe the images in a natural way.

ACKNOWLEDGMENTS

Tao Wang is funded by China Scholarship Council, and we gratefully acknowledge the support of DFG SFB/TR8 I5 and NSFC 61272336.

References

Allen, James F. 1983. 'Maintaining Knowledge About Temporal Intervals'. *Commun. ACM* 26, no. 11: 832–843. http://doi.acm.org/10.1145/182.358434.

Bay, Herbert, Ess, Andreas, Tuytelaars, Tinne & Gool, Luc Van. 2008. 'Speeded-Up Robust Features (SURF)'. *Computer Vision and Image Understanding* 110, no. 3: 346 – 359. http://www.sciencedirect.com/science/article/pii/S1077314207001555. Similarity Matching in Computer Vision and Multimedia.

Coyne, Bob, Bauer, Daniel & Rambow, Owen. 2011. 'Vignet: Grounding language in graphics using frame semantics'. In 'Proceedings of the ACL 2011 Workshop on Relational Models of Semantics', 28–36. Association for Computational Linguistics.

Dalal, N. & Triggs, B. 2005. 'Histograms of oriented gradients for human detection'. In 'Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on', vol. 1, 886–893 vol. 1.

Escalante, Hugo Jair, Hernández, Carlos A., Gonzalez, Jesus A., López-López, A., Montes, Manuel, Morales, Eduardo F., Sucar, L. Enrique, Villaseñor, Luis & Grubinger, Michael. 2010. 'The segmented and annotated {IAPR} TC-12 benchmark'. *Computer Vision and Image Understanding* 114, no. 4: 419 – 428. http://www.sciencedirect.com/science/article/pii/S1077314209000575. Special issue on Image and Video Retrieval Evaluation.

Goyal, Roop K. & Egenhofer, Max J. 2000. 'Consistent Queries over Cardinal Directions across Different Levels of Detail'. In 'Proceedings of the 11th International Workshop on Database and Expert System Applications', 876–880.

Li, Sanjiang. 2013. 'Cardinal Directions between Regions: A Comparison of Two Models'. Research article.

Lieto, Antonio. 2014. 'A Computational Framework for Concept Representation in Cognitive Systems and Architectures: Concepts as Heterogeneous Proxytypes'. *Procedia Computer Science* 41: 6 – 14. http://www.sciencedirect.com/science/article/pii/S1877050914015233. 5th Annual International Conference on Biologically Inspired Cognitive Architectures, 2014 {BICA}.

Liu, Weiming, Zhang, Xiaotong, Li, Sanjiang & Ying, Mingsheng. 2010. 'Reasoning about cardinal directions between extended objects'. *Artificial Intelligence* 174, no. 12 - 13: 951 – 983. http://www.sciencedirect.com/science/article/pii/S0004370210000834.

Lowe, D.G. 1999. 'Object recognition from local scale-invariant features'. In 'Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on', vol. 2, 1150–1157 vol.2.

Schneider, Markus, Chen, Tao, Viswanathan, Ganesh & Yuan, Wenjie. 2012. 'Cardinal Directions Between Complex Regions'. *ACM Trans. Database Syst.* 37, no. 2: 8:1–8:40. http://doi.acm.org/10.1145/2188349.2188350.

Socher, Richard, Karpathy, Andrej, Le, V. Quoc, Manning, D. Christopher & Ng, Y. Andrew. 2014. 'Grounded Compositional Semantics for Finding and Describing Images with Sentences'. *Transactions of the Association of Computational Linguistics – Volume 2, Issue 1* 207–218. http://aclweb.org/anthology/Q14-1017.

Wang, Li & He, Dong-Chen. 1990. 'Texture Classification Using Texture Spectrum'. *Pattern Recogn.* 23, no. 8: 905–910. http://dx.doi.org/10.1016/0031-3203(90)90135-8.

Zhang, Dengsheng, Islam, Md. Monirul & Lu, Guojun. 2012. 'A review on automatic image annotation techniques'. *Pattern Recognition* 45, no. 1: 346 – 362. http://www.sciencedirect.com/science/article/pii/S0031320311002391.