TOWARDS AN IMAGE UNDERSTANDING SYSTEM FOR MULTIPLE VIEWPOINTS

PAULO E. SANTOS*, DANILO NUNES DOS SANTOS†

*Centro Universitario da FEI,
S. Paulo, Brazil

Emails: psantos@fei.edu.br, nunesdanilo@gmail.com

Abstract— In this paper we present preliminary results towards the construction of a system capable of executing image understanding from multiple viewpoints from the standpoint of qualitative spatial reasoning and feature matching.

Keywords— Qualitative Spatial Reasoning, Cognitive Robotics, View Combination, Artificial Intelligence

Resumo— Neste artigo apresenta-se resultados preliminares sobre o desenvolvimento de um sistema capaz de executar interpretacao de imagens provenientes de multiplos pontos de vista utilizando conceitos de raciocinio espacial qualitativo e *feature matching.*

Palavras-chave— Qualitative Spatial Reasoning, Cognitive Robotics, View Combination.

1 Introduction

The modern research on scene interpretation is based on the development of probabilistic methods motivated by the need to deal with sensor noisy and image uncertainty (Lavee et al., 2009). Probabilistic methods, however, are propositional, imposing restrictions in their capability to represent general domain knowledge and their applicability on problems containing a possibly unbounded number of objects. Logic-based image interpretation, on the other hand, tackles exactly the problem of the effective representation of general facts about the domain, as well as the generalisation of these facts to problems with infinite variables (Santos, 2010). Thus, research on logic-based image interpretation does not preclude the use of probabilistic methods, but complements them by making explicit the knowledge content of a domain.

The first framework for a logic-based scene interpretation system was proposed in (Reiter and Mackworth, 1989) where three sets of axioms were defined to constrain the number of possible interpretations of the scenes observed. Therefore, the scene interpretation process is reduced to a constraint satisfaction problem. The SIGMA system (Matsuyama and Hwang, 1990) successfully deploys the ideas proposed by Reiter and Mackworth on the field of aerial image interpretation. Some properties of the formalism used by the SIGMA system were further developed in (Schroeder and Neumann, 1996) and recently revisited and incorporated into a description logic setting (Neumann and Möller, 2007).

In this context, our contributions follows the ideas proposed in (Shanahan, 1996), where a logical formalism is developed to rigorously define the information obtained from a robot's sensors in terms of symbols hypothesising the existence, location and shape of the observed objects. In previous works (Santos and Shanahan, 2002; Santos, 2007; dos Santos et al., 2009), we have developed a theory aimed at the automatic scene understanding from a robot's viewpoint. In particular, (Santos, 2007) presents a formalism capable to interpret events such as *approaching*, *receding*, or *coalescing* from pairs of subsequent images obtained by a mobile robot's stereopair. In order to further interpret these image-related events, we developed an abductive procedure for hypothesising on the possible changes that might have occurred with the domain objects that could explain the image events.

In another work (dos Santos et al., 2009), we developed a framework capable of interpreting events based on an arbitrary long image sequence. In this case, events such as the *rotation* of an object around a reference point; one object *following* another; and, one object *trespassing* another were formally defined within a dynamic logic and further used in a scene interpretation procedure involving perceptually indistinguishable objects. This system, however, could not handle uncertainty in the scenes. In (Fenelon et al., 2010; Santos, Hummel, Fenelon and Cozman, 2010) we developed a spatial reasoning system in probabilistic logics that was applied on the task of interpreting images from traffic scenes from the viewpoint of a camera at the driver's position (Souza and Santos, 2011). Analogously, we applied a probabilistic logic system to induce general facts from complex videos obtained by a number of surveillance cameras in an airport apron (Dubba et al., 2011).

The development of cognitive vision systems for a mobile robot was also considered in our previous work. Based on recent psychological findings (summarised in (Dee and Santos, 2011)), we formalised the perception of shadows in terms of a qualitative spatial theory and used in to interpret the robot's environment (Santos et al., 2009; Santos, Dee and Fenelon, 2010; Santos et al., 2011).

In all of these works, however, the scenes were observed from the viewpoint of a single robot, that has only its own knowledge base available for the scene interpretation process. This paper presents the initial steps towards extending our previous logic-based image interpretation systems towards the interpretation of scenes as observed by multiple, distinct, viewpoints. To this end, the next section presents some ideas on extending current qualitative spatial reasoning systems towards 3D formalisms capable of representing the relative position of multiple viewpoints. Section 3 presents preliminary results on extending feature matching algorithms for handling scenes observed from *extreme* viewpoints. Sections 4 and 5 present some discussions on the results described, concluding this paper.

2 Qualitative Spatial Reasoning about Viewpoints

In this paper a viewpoint (represented by lowercase Greek letters) is a unary vector in R^3, whose origin represents the viewpoint location and its direction represents the direction of the robot's front. The direction is basically given by the angle of the robot's front with respect to the magnetic Earth's North, as given by the robot's compass and inertial measurement unit (IMU).

2.1 Viewpoint-relative position

The first step in our characterisation of a domain with multiple viewpoints is to define the means to specify the relative position of objects with respect to a viewpoint (including the relative position of one viewpoint with respect to another). For that we define an egocentric 3D qualitative reference system that recalls the idea of Celestial Sphere, as shown in Fig. 1. Whereas the viewpoint is the origin of a sphere that represents a discretisation of the space around it into altitude and longitude lines (dashed circles in Fig. 1). In a nutshell the space around the viewpoint ν is discretised into the *altitude-longitude* categories *at*, *near*, far, *upper*, *lower*, *below* and *under*. It is worth pointing out that these categories are independent of the viewpoint direction, i.e., the normal of the top-half sphere is always pointing towards the sky, and the limit between upper and lower are always parallel to the horizontal line. The extent of each of these spatial concepts depends on the application domain (which may include priorities on the tasks to be accomplished, the remaining battery charge and the robot's capabilities).

Formally, the categories *at*, *near*, far, *upper*, *lower*, *below* and *under* are defined imposing appropriate thresholds on the distances from the viewpoint at the origin of the sphere. Intuitively, *at* represents the viewpoint's local neighbourhood, which in the case of a ground robot can be defined as the smallest sphere enclosing its occupancy region, and in the case of the quadrotor, it can be defined as a safety distance that could be used to avoid collisions between other agents; *near* represents a region of space that can be reached in a few minutes by the robot; the intuition of far is analogous. The altitude categories *upper*, *lower*, *below* and *under* are fixed in length and represent the "near" and "far" portions of space around a viewpoint on the altitude dimension. It is worth pointing out that, on a complete planar field, the personal sphere related to an any-terrain robot will be in fact a half sphere (only the upper half of Fig. 1).

Taking into consideration the viewpoint's direction, the space around ν can also be discretised into the *relative positions*: $left$, $right$, $front$, *back*, $leftfront$, $rightfront$, $leftback$ and *rightback* using a version of the *8-Star Calculus* (Renz and Mitra, 2004) as shown in Fig. 2. Combining the *altitude-longitude* categories with the *relative positions* we have a very rich qualitative discretisation of the space around a viewpoint. Fig. 3 shows a cut at the equator of the personal sphere making explicit this discretisation, whereby we use the following abbreviation f, b, r, l, lf, lb, rf, rb representing respectively $front$, *back*, $right$, $left$, $leftfront$, $leftback$, $rightfront$ and *rightback*.

In this work the relative positions and the altitude-longitude categories described above are binary relations between viewpoints and are defined according to the Euclidean distances between the robots (whose 3D locations are given by the robots' GPS) and the directions of the robots' gaze (as given by the robot's IMU). This set of relations can be understood as a 3D version of the Ternary Point Configuration Calculus (Moratz and Ragni, 2008), whereby the origin is a viewpoint and the *relatum* is given by the direction of the robot's gaze with respect to its compass (i.e. it is a point at the infinity). A formal definition of the relations presented above is outside of the scope of this paper and is presented elsewhere (Santos, 2013).

The purpose of defining this discretisation of the local space around a viewpoint is two fold: first, in some situations in field robotics applications the exact position of the robots is sometimes irrelevant: what is important is to keep the robot in a determinate region of space with respect to a natural reference point. Second, a key issue in a mixed initiative system is the interaction with humans. In particular, search and rescue situations usually include human volunteers who are not acquainted with robotics methods and mathematical functions, in general. Therefore the use of (a portion of) natural language to describe the actions and perceptions of both robotic and human agents is an essential feature.

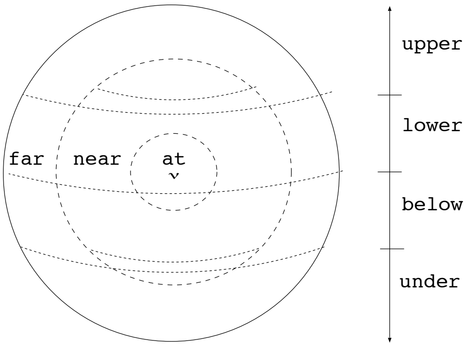


Figure 1: The personal sphere wrt viewpoint ν.

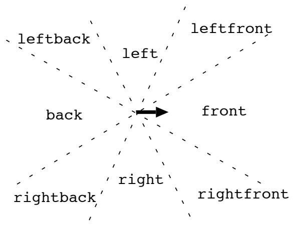


Figure 2: relative positions from a viewpoint.

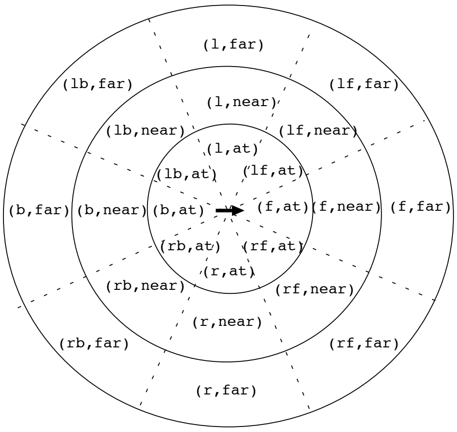
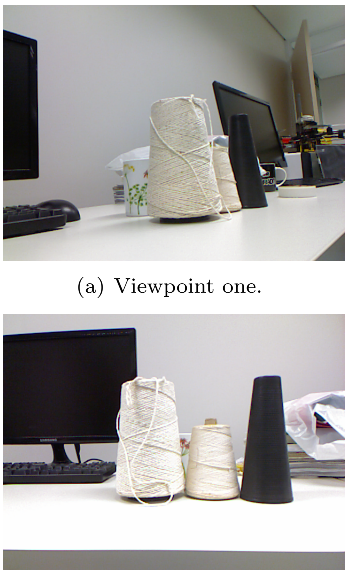


Figure 3: Qualitative discretisation at the equator of the personal sphere.

The formal characterisation of relative position between viewpoints presented above comes alongside to the development of new feature matching algorithms capable of finding a match between images seem from *extreme* viewpoints, such as from a global viewpoint (picked out by a quadrotor, for instance) and a ground robot observing the same scene. Preliminary results on the development of feature matching algorithms of this nature are described in the next section.

3 Understanding Multiple Viewpoints

One important goal of this work is the combination of the knowledge obtained from multiple (distinct) viewpoints. Starting from distinct viewpoints regarding the same scene, the task is to identify common features, specially in poor conditions as partial occluded objects. In Figures 4(a) and 4(b) it is possible to see that both images refer to the same environment, although from different viewpoints. Standard feature-matching algorithms fail in this case due to the high variation in these viewpoints. It is also important to notice that the *monitor screen* is partially hidden in figure 4(a), so it leads to difficulties in the matching process.



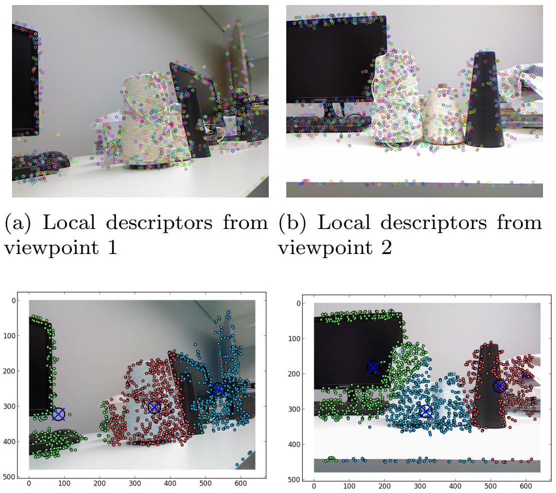
(a) Viewpoint one.

(b) Viewpoint two

Figure 4: Snapshots from distinct viewpoints of the same scene.

This work proposes a new approach to perform feature matching. In the method developed here, first of all, it is necessary to extract local descriptors by a common feature extractor like SURF (Bay et al., 2008) (Figs. 4(a) and 5(b)). Then, before performing feature matching against each keypoint, as usually done, our method clusters the descriptors. In order to improve the results, the dimension was increased considering also the spatial position (x, y) and color (r, g, b) of the descriptors. The clusters thus generated applying these ideas on the images in Figures 4(a) and 4(b) are shown in Figures 5(c) and 5(d), respectively.

Next, for each element in the source cluster the method performs a search in the target cluster. If the distance between descriptors of the source and target clusters is below a defined threshold, the pair is considered as a match. The distance from clusters is calculated considering the relation between the number of individual element



(a) Local descriptors from viewpoint 1 (b) Local descriptors from viewpoint 2

(c) Viewpoint one. (d) Viewpoint two

Figure 5: Clusters from descriptors.

matches and the average distance between them. This process takes in account both directions, it means, cluster A in source image is considered similar to cluster B in target image, if and only if both the distances from A to B and B to A are the smallest, otherwise it is discarded. Figure 6 shows the matched regions with respect to the input images on Figure 4.

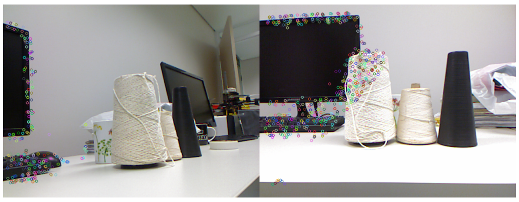


Figure 6: Most similar regions according to clusters from figs. 5(c) 5(d).

4 Discussion

In the results obtained it is possible to observe that the monitor screen was roughly matched in both images. Although, it worked as expected, i.e., found similarities between the same object from different viewpoints. When considering only individual matches, we realised that the points matched did not work properly as most of the pairs matched happened between distinct parts of the object on one viewpoint with respect to the same object seen on another. This effect was due to overweighted neighbourhood around each descriptor.

An extension of this approach considering global information, or even the spatial relations between local descriptors inside a cluster, could improve the method and produce more reliable results.

5 Conclusion

This paper described ongoing research on extending the existing spatial reasoning systems towards the explicit consideration of multiple viewpoints of a scene within a single (combined) formalism. We also presented preliminary results on the development of a pattern matching algorithm capable of identifying objects on scenes observed from distinct (extreme) viewpoints.

Acknowledgments

This work receives financial support from FAPESP, proc. 2012/04089-3. The first author is partially supported by CNPq - PQ2.

The second author is partially supported by CAPES under PROSUP/CAPES program proc. 00.513102-4.

References

Bay, H., Ess, A., Tuytelaars, T. and Van Gool, L. (2008). Speeded-up robust features (surf), *Comput. Vis. Image Underst.* **110**(3): 346–359.

Dee, H. M. and Santos, P. E. (2011). The perception and content of cast shadows: An interdisciplinary review, *Spatial Cognition & Computation* **11**(3): 226–253.

dos Santos, M., de Brito, R. C., Park, H.-H. and Santos, P. (2009). Logic-based interpretation of geometrically observable changes occurring in dynamic scenes, *Applied Intelligence* **31**(2): 161–179.

Dubba, K., Santos, P. E., Cohn, A. and Hogg, D. (2011). Probabilistic relational learning of event models from video, *21st International Conference on Inductive Logic Programming (ILP 2011)*.

Fenelon, V., Santos, P., Hummel, B. and Cozman, F. (2010). Encoding spatial domains with relational bayesian networks, *Spatio-temporal Dynamics Workshop*, pp. 49–54.

Lavee, G., Rivlin, E. and Rudzsky, M. (2009). Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video, *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* **39**(5): 489 –504.

Matsuyama, T. and Hwang, V. S. (1990). *SIGMA: A Knowledge-Based Image Understanding System*, Plenum Press, New York, U.S.

Moratz, R. and Ragni, M. (2008). Qualitative spatial reasoning about relative point position, *J. Vis. Lang. Comput.* **19**(1): 75–98.

Neumann, B. and Möller, R. (2007). On scene interpretation with description logics, *Image and Vision Computing, Special Issue on Cognitive Vision* .

Reiter, R. and Mackworth, A. (1989). A logical framework for depiction and image interpretation, *Artificial Intelligence* **41**(2): 125–155.

Renz, J. and Mitra, D. (2004). Qualitative direction calculi with arbitrary granularity, *In Proceedings of the 8th Pacific Rim International Conference on Artificial Intelligence*, Springer, pp. 65–74.

Santos, P. (2007). Reasoning about depth and motion from an observer's viewpoint, *Spatial Cognition and Computation* **7**(2): 133–178.

Santos, P. (2013). Viewpoints in evidence. Manuscript.

Santos, P., Dee, H. M. and Fenelon, V. (2009). Qualitative robot localisation using information from cast shadows, *IEEE International Conference on Robotics and Automation*, IEEE, pp. 220–225.

Santos, P., Dee, H. M. and Fenelon, V. (2010). Knowledge-based adaptive thresholding from shadows, *Proceeding of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, IOS Press, Amsterdam, The Netherlands, The Netherlands, pp. 1089–1090.

Santos, P. E. (2010). *Tutoriais do XVIII Congresso Brasileiro de Automatica*, Vol. 1, Cultura Academica, chapter Raciocinio e percepcao espacial: uma abordagem logica, pp. 127–147.

Santos, P. E., Fenelon, V. and Dee, H. (2011). Reasoning about shadows in a mobile robot environment. under submission.

Santos, P. E., Hummel, B., Fenelon, V. and Cozman, F. (2010). Probabilistic encoding of spatial domains, *Proc of the first International Workshop on Uncertainty in Description Logics*, pp. 1–6.

Santos, P. and Shanahan, M. (2002). Hypothesising object relations from image transitions, *in* F. van Harmelen (ed.), *Proc. of ECAI*, Lyon, France, pp. 292–296.

Schroeder, C. and Neumann, B. (1996). On the logics of image interpretation: Model construction in a formal knowledge representation framework, *International Conference on Image Processing*, Vol. 2, Switzerland, pp. 785–788.

Shanahan, M. (1996). Robotics and the common sense informatic situation, *Proc. of ECAI*, Budapest, Hungary, pp. 684–688.

Souza, C. R. and Santos, P. E. (2011). Probabilistic logic reasoning about traffic scenes, *The 12th Conference Towards Autonomous Robotic Systems (TAROS)*, Springer-Verlag, Berlin, Heidelberg, pp. 219–230.