Interest-Point-Based Landmark Computation for Agents' Spatial Description Coordination

J. I. Olszewska
School of Computing and Technology, University of Gloucestershire, The Park, Cheltenham, GL50 2RH, U.K.

Keywords: Qualitative Spatial Reasoning, Object Detection, Local Feature Descriptors, Feature Extraction, Visual Scene Understanding, Automated Image Annotation, Robotics, Autonomic Agents.

Abstract: In applications involving multiple conversational agents, each of these agents has its own view of a visual scene, and thus all the agents must establish common visual landmarks in order to coordinate their space understanding and to coherently share generated spatial descriptions of this scene. Whereas natural language processing approaches contribute to define the common ground through dialogues between these agents, we propose in this paper a computer-vision system to determine the object of reference for both agents efficiently and automatically. Our approach consists in processing each agent's view by computing the related, visual interest points, and then by matching them in order to extract the salient and meaningful landmark. Our approach has been successfully tested on real-world data, and its performance and design allow its use for embedded robotic system communication.

1 INTRODUCTION

Communication between agents about space is of prime importance for actions requiring spatial coordination of intelligent agents such as robots operating in rescue activities or in assistive aid.

In these joint actions, the conversational agents, i.e. the speaker and the hearer, should reach a shared understanding of the scene they observe and/or evolve in. For this purpose, they usually acquire knowledge about this scene as well as its objects (Olszewska, 2011), and generate qualitative spatial descriptions of the scene, by using semantic concepts such as "to the right", "at two oclock" (Olszewska and McCluskey, 2011) or "above" (Olszewska, 2013). However, notions involving spatial relations require the definition of a reference object in the scene. Hence, the agents must adopt a common ground in order to coordinate their spatial descriptions of the scene (Ma et al., 2012).

Such common reference could be of different nature (Anacta et al., 2014) in a discourse describing and interpreting visual scenes such as presented in Fig. 1. Indeed, the reference could be defined by some third object, leading to a relative reference (e.g. the palm tree), distinct form the reference and related objects (e.g. the speaker and the hearer, respectively). The reference could also be the object itself, i.e. intrinsic reference (e.g. the coffee table). On the other hand, the reference could refers to some global reference point, set as an extrinsic reference (e.g. the North). Hence, the scene could be grounded by inferred or explicit reference to these specific objects (Levinson, 2003). For example, the sofa may be"to the right of the coffee table", "behind the coffee table", or "North of the coffee table". This leads to a situated conversation as the agents have now a visual common ground, while objects of the scene may be described from different perspectives corresponding to each agent's view of the 3D scene (Olszewska, 2015a).

In Natural Language Processing, methods are usually setting visual landmarks through dialogue between agents (Jurafsky and Martin, 2000). However, this latter step is a limitation for effectively coordinating spatial descriptions between dialogue partners, in case one agent is human and the other one is a non-human agent as for applications involving robots (Summers-Stay et al., 2014).

Cognitive approaches such as (Zhanga et al., 2014) focus on the analysis of the visual scene in order to identify objects which attract human agents' attention, leading to a fast definition of reference objects. This quantitative study highlights that selecting as a reference object salient objects visible in all agents' views could positively impact on the land-

Olszewska, J.
Interest-Point-Based Landmark Computation for Agents' Spatial Description Coordination.
DOI: 10.5220/0005847705660569
In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART 2016) - Volume 2*, pages 566-569
ISBN: 978-989-758-172-4

(a)

(b)

Figure 1: CANDELA dataset image of the studied visual scene observed from (a) a speaker's viewpoint; (b) a hearer's viewpoint.

mark definition effectiveness, while works such as (Watson et al., 2004) demonstrate that intrinsic references are the most widely adopted type of references among agents.

On the other hand, in Computer Vision, methods such as presented in (Alsuqayhi and Olszewska, 2013), (Bhat and Olszewska, 2014), (Olszewska, 2015b), have proven to be efficient for automatic annotation of visual scenes and for automated scene understanding.

Hence, in this work, we propose to use a computer-vision approach to automatically identify common visual landmarks in order to allow the automated coordination of spatial descriptions between any type of agent, i.e. human/non-human one.

Our method does not imply constraints on the geometrical properties of visual landmarks (see Fig. 3).

The contribution of this paper thus consists in the automatic definition of visual landmarks by processing the visual feature descriptors of agents' different views of the same scene in order to extract common-ground, reference objects between these two agents.

The paper is structured as follows. In Section 2, we present our approach to compute visual landmarks in an automated way. The performance of our computer-vision approach successfully tested for different types of agents in real-world, indoor situations are reported and discussed in Section 3. Conclusions are drawn up in Section 4.

2 PROPOSED APPROACH

To automatically extract the common landmark for coordinating spatial descriptions of multi-view scenes, we propose the following approach involving computer-vision based techniques (see Fig. 2) as explained in Sections 2.1-2.2.

2.1 Detecting Objects of Interest

Firstly, each view is processed separately in order to extract the visual information. Hence, an interest point detector (Alqaisi et al., 2012) is applied to the speaker's and hearer's views such as Figs. 1(a) and (b), respectively.

Once the interest points have been detected in each of the view of the scene, a candidate landmark object (e.g. the coffee table) is matched with each view to determine the potential objects of reference, i.e. the objects of interests, as illustrated in Fig. 3(a) and Fig. 3(b), respectively. This process is involving the automatic labeling method as described in (Olszewska, 2012).

2.2 Computing Visual Landmarks

Next, the objects which have been detected in the different views are matched with each other in order to set the common the landmark as displayed in Fig. 3(c).

All the matching computations we performed rely on the computation of the Hausdorff distance $d_H(A,B)$ defined as follows:

$$d_H(A,B) = max\Big(d_h(A,B), d_h(B,A)\Big), \qquad (1)$$

where $d_h(A,B)$ is the directed Hausdorff distance from A to B defined as

$$d_h(A,B) = \max_{a\in A}\min_{b\in B} d_P(a,b), \qquad (2)$$

with $d_P(a,b)$, the Minkowski-form distance based on the L_P norm, and defined as

$$d_P(a,b) = \left(\sum_k (a_k - b_k)^P\right)^{1/P}, \qquad (3)$$

and involve the embedded matching algorithm (Algorithm 1) (Alqaisi et al., 2012) where A and B are the two finite sets of SIFT local descriptors, detected in

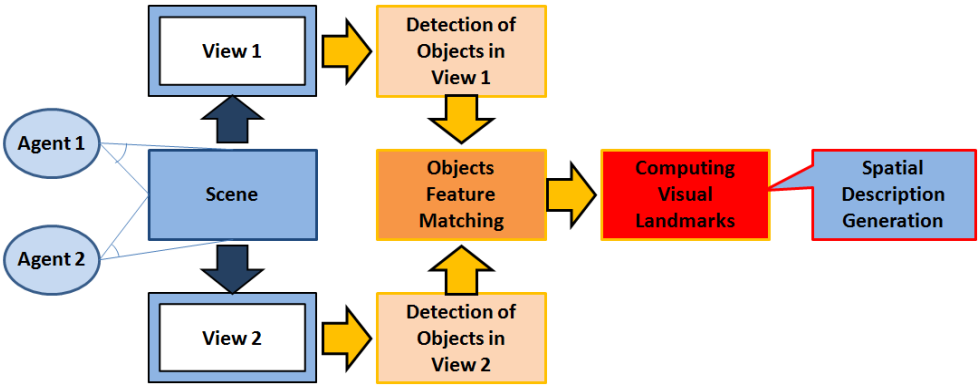


Figure 2: Architecture of the proposed automated computation of visual landmarks for coordinating spatial descriptions in discourse between agents.

Algorithm 1: Embedded Matching.

Given $A' = A, B' = B, M = \varnothing$,
for all $a_i \in A'$ **do**
  for all $b_j \in B'$ **do**
    repeat
      if

$$\begin{aligned} & d_P(a_i,b_j) = \min_{b \in B'} d_P(a_i,b) \\ \wedge \quad & d_P(b_j,a_i) = \min_{a \in A'} d_P(b_j,a) \\ \wedge \quad & d_P(a_i,b_j) \leq d_H(A,B) \\ \wedge \quad & d_P(b_j,a_i) \leq d_H(A,B) \end{aligned}$$

      then

$$\begin{aligned} & (a_i,b_j) \subset M \\ \wedge \quad & A' = A \backslash \{a_i\} \wedge B' = B \backslash \{b_j\} \end{aligned}$$

      end if
    until $A' \neq \varnothing \vee B' \neq \varnothing$
  end for
end for
return M

each view and and where M is the doubly matched feature set.

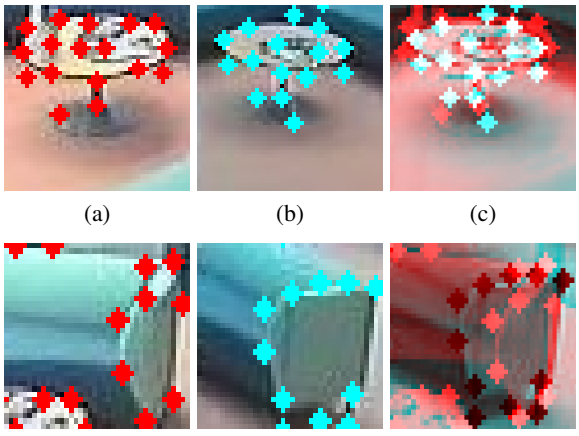
The object similarity measure $d_S(A,B)$ is then defined as follows

$$d_S(A,B) = \frac{\#M}{\frac{\#A+\#B}{2}}, \tag{4}$$

3 EXPERIMENTS

We carried out two types of experiments to validate our system. In the first experiment, we used a publicly-available dataset called CANDELA which contains images showing the same indoor scene captured from different point of views (Fig. 1) by two different cameras. This configuration maps the set-up consisting of two cameras, each connected to a MatLab-equipped PC, with one modeling the speaker agent and the other one the hearer agent. Processing the image sent by the speaker and the one acquired by the hearer leads to compute common landmarks based on matching the detected interest points in each view with a candidate object of reference and with each other. To test our approach in this case, matching has been performed repeatedly on different views of the dataset and for different objects of interest (Fig. 3). It is worth noting that changes in landmark objects' poses due to the different views have a major impact on this matching process. Experimental results have been compared with ground truth data obtained by two human agents looking each at a different view. Our automatic system shows excellent performance, achieving 94% of accuracy, and is thus promising to be embedded in an autonomous process.

In the second experiment, the set-up is composed of a speaker agent modeled by a camera Arducam acquiring the view 1 pinned to an Arduino board connected to a Bluetooth TTL transceiver module. The hearer agent consists of a webcam recording the view 2 and linked to a PC running MatLab software processing the images of the different views and connected with a Bluetooth master module in order to communicate with the agent 1. As long as the Bluetooth connection is operating properly, the accuracy of the system setting the common landmark is of 92% as assessed by comparing the matching results with two humans agents looking at each of the views recorded by the different cameras. These results are excellent compared to other approaches such as (Summers-Stay et al., 2014), and could be further improved by adding some image pre-processing techniques to cope with light variations in the cap-



(a) (b) (c)

(d) (e) (f)

Figure 3: Computing visual landmark: (1st column) detecting interest points in the first view; (2nd column) detecting interest points in the second view; (3rd column) matching both views' interest points.

tured images, especially those acquired by the mobile speaker agent compared to the static hearer agent.

4 CONCLUSIONS

This paper presents an automatic and accurate method to objectively define common visual landmarks in order to coordinate spatial descriptions generated by different agents, each with a different view of the same scene. Hence, the common ground is computed by detecting interest points in all the agents' views and by applying the Hausdorff-enhanced matching of these points in order to extract the common salient object visible in both agents' views. Our approach is a new application of computer-vision local feature descriptor computation in context on agent communication systems. This new automated process could be successfully integrated in robotic applications as demonstrated.

REFERENCES

Alqaisi, T., Gledhill, D., and Olszewska, J. I. (2012). Embedded double matching of local descriptors for a fast automatic recognition of real-world objects. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'12)*, pages 2385–2388.

Alsuqayhi, A. and Olszewska, J. I. (2013). Efficient optical character recognition system for automatic soccer player's identification. In *Proceedings of the IAPR International Conference on Computer Analysis of Images and Patterns Workshop (CAIP'13)*, pages 139–150.

Anacta, V. J. A., Schwering, A., and Li, R. (2014). Determining hierarchy of landmarks in spatial descriptions. In *Proceedings of the International Conference on Geographic Information Science (GIScience'14)*.

Bhat, M. and Olszewska, J. I. (2014). DALES: Automated tool for detection, annotation, labelling and segmentation of multiple objects in multi-camera video streams. In *Proceedings of the ACL International Conference on Computational Linguistics (COLING'14)*, pages 87–94.

Jurafsky, D. and Martin, J. H. (2000). *Dialogue and conversational agents*, chapter 19, pages 719–761. Prentice Hall.

Levinson, S. C. (2003). *Space in Language and Cognition: Explorations in Cognitive Diversity*, chapter 5. Cambridge Press University.

Ma, Y., Raux, A., Ramachandran, D., and Gupta, R. (2012). Landmark-based location belief tracking in a spoken dialog system. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'12)*, pages 169–178.

Olszewska, J. I. (2011). Spatio-Temporal Visual Ontology. In *Proceedings of the 1st EPSRC/BMVA Workshop on Vision and Language (VL'11)*.

Olszewska, J. I. (2012). A new approach for automatic object labeling. In *Proceedings of the 2nd EPSRC/BMVA Workshop on Vision and Language (VL'12)*.

Olszewska, J. I. (2013). Clock-modeled ternary spatial relations for visual scene analysis. In *Proceedings of the ACL International Conference on Computational Semantics Workshop*, pages 20–30.

Olszewska, J. I. (2015a). 3D Spatial reasoning using the clock model. *Research and Development in Intelligent Systems XXXII, Springer*, pages 147–154.

Olszewska, J. I. (2015b). "Where is my cup?" - Fully automatic detection and recognition of textureless objects in real-world images. *Lectures Notes in Computer Science, Springer*, 9256:501–512.

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

Summers-Stay, D., Cassidy, T., and Voss, C. R. (2014). Joint navigation in commander/robot teams: Dialog and task performance when vision is bandwidth-limited. In *Proceedings of the ACL International Conference on Computational Linguistics*, pages 9–16.

Watson, M. E., Pickering, M. J., and Branigan, H. P. (2004). Alignment of reference frames in dialogue. In *Proceedings of the Annual Conference of the Cognitive Science Society*.

Zhanga, X., Lia, Q.-Q., Fang, Z.-X., Lu, S.-W., and Shaw, S.-L. (2014). An assessment method for landmark recognition time in real scenes. *Journal of Environmental Psychology*, 40:206–217.