Image and Vision Computing 34 (2015) 42–44

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

The Cognitive Dialogue: A new model for vision implementing common sense reasoning☆

Yiannis Aloimonos, Cornelia Fermüller

Computer Vision Laboratory, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received 15 October 2014
Accepted 28 October 2014
Available online 6 December 2014

Keywords:
Vision and language
Integration of perception, action, and cognition
Cognition modulating image processing

ABSTRACT

We propose a new model for vision, where vision is part of an intelligent system that reasons. To achieve this we need to integrate perceptual processing with computational reasoning and linguistics. In this paper we present the basics of this formalism.

1. Perception as recovery

The goal of classical Computer Vision has been to create 3d descriptions of the scene or recognize the scene by assigning labels to objects and actions. The labels would be given to symbolic reasoning systems, of the kind Artificial Intelligence develops, that would reason about the world. This separation of vision from cognitive considerations allowed for a specialized and formal analysis of images, and allowed Computer Vision to become a discipline on its own. Even famous psychologists subscribing to the paradigm wrote extensively on the cognitive impenetrability of visual perception, implying that the workings of visual perception are shielded from any cognition and presenting visual perception as a mechanistic black box that delivers labels through recognition.

Current practice has suggested repeatedly that going from “pixels” to “symbols” in a bottom up manner is very hard, if not impossible. It appears that knowledge of some form comes into the process quite early. In the classical framework, there is only one place where perception and cognition meet. This does not seem to fit well with our intuitive understanding of perception and thinking. Human behavior is active and exploratory. We continuously shift our gaze to different locations in the scene in front of us. We recognize objects, sounds and actions and this leads us to fixate again at a new location, and so on. Humans interpret perceptual input by using their knowledge of images, sounds, actions and objects, along with the perceptual operators that extract information from signals.

2. Perception as inference

Theorists of perception have long understood that signal analysis is not enough to produce an understanding of the scene; there must be some additional source of information beyond retinal images that is used in the process of seeing. The famous visual psychologist von Helmholtz proposed that the additional knowledge is brought in through a process of inference. Since we are not aware of it, he called it *unconscious inference*. While most of the literature discusses this inference only with respect to geometrical and physical constraints in the world, we argue that the idea can be taken further. Adding any form of knowledge to the signal processing can be implemented as inference or reasoning. The prior knowledge can be unconscious about the physics of the world and conscious about likely configurations of objects, events and their spatio-temporal relations. We conclude that perception interacts continuously with cognition at different levels of abstraction: to guide attention, to make predictions, to constrain the search space for recognition, and to reason over what is being perceived. This is an interactive bottom-up and top-down interaction; as information is exchanged between vision and cognition, meaning emerges.

3. The use of language

At the technical level, infusing perception with reasoning can happen through knowledge-based engineering and the use of language. There has been a recent interest in research in the field of Computer Vision to introduce additional higher-level knowledge about image relationships into the interpretation process (e.g. [3,5,6]). While current studies get this additional information from captions or accompanying text, more advanced language processing can be used to obtain additional high level information.

☆ This paper has been recommended for acceptance by Sinisa Todorovic.

E-mail addresses: yiannis@cs.umd.edu (Y. Aloimonos), fer@cfar.umd.edu (C. Fermüller).

http://dx.doi.org/10.1016/j.imavis.2014.10.010
0262-8856/© 2014 Elsevier B.V. All rights reserved.

Linguists and computational linguists have a longstanding interest in modeling lexical semantics, i.e. conceptual meanings of lexical items and how these lexical items relate to each other [1] and have created resources where information about different concepts, such as cause–effect, performs-functions, used-for, and motivated-by, can be obtained. For example, the Word-Net database relates words through synonymy (words having the same meaning, like "argue" and "contend") and hypernymy ("is–a" relationships, as between "car" and "vehicle"), among many others [8]. Linguistics also has created large text corpora and statistical tools so we can obtain probability distributions for the co-occurrence of any two words, such as how likely a certain noun co-occurs with a certain verb.

Using these linguistic tools, how can we aid vision to build better systems for interpreting images? One way is to use linguistic information as a *contextual system* that provides additional information to the interpretation, as already utilized in some multimedia systems. For example, certain objects are likely to co-occur, such as "tables" often co-occur with "silverware" and "glasses". But if we consider vision an active process, there is more. Let's say you are in a kitchen. Because you have prior knowledge about kitchens, their structure and the actions taking place in them and a large part of this knowledge is expressed in language, we can utilize this information during visual inspection. A knife in the kitchen will most probably be used for "cutting" a food item, so the vision can look for it. Then language acts as a *high level prior knowledge system* that aids perception. Or, let's say you observe someone pick up the knife and put in the drawer. Given this, you know that the object is not gone, but just hidden from sight. In this case, language acts as *part of a reasoning process*. When humans interpret a visual scene, we fixate at some location and recognize nouns, verbs, adjectives, adverbs and prepositions. Because the linguistic system is highly structured, these recognitions produce a large number of inferences about what could be happening in the scene. This leads us to fixate at a new location, and the same process is repeated. In other words, language acts as *part of an attention mechanism*.

4. The Cognitive Dialogue

In a real sense, during the process of vision, perceptual processes are interacting with language processes and motor actions. Our attention is guided by low level perceptual object features and/or movements, but also by high level knowledge and our overall goals. Object and action recognition themselves interact continuously with prior knowledge that formulates expectations and constrains the recognition search space. Reasoning is used to make sense out of visual input correcting visual recognition at times toward solutions that make sense in one's context. This is a dynamic interaction of cognitive processes that is generally agreed upon but has not been implemented yet computationally. We suggest that this interaction should be implemented as a dialogue computationally to achieve scalable visual scene analysis by intelligent systems (e.g. in a message passing framework). To give an example, let's say the goal is the production of a semantic description of the scene in view. The way this can be achieved, is by having the language processes (LP) and the visual processes (VP) engage in some form of a Cognitive Dialogue, in language. The LP can ask the VP a number of questions, such as: is there <noun>? in the scene? Where is it? What is next to <noun>? Where did the agent that performed <action> go afterwards?. By allowing the LP to ask questions and receive answers, and repeat the process, we bring forward the whole power of language in the semantic analysis, something that was not possible before. If we also include the motor processes, MP, and the auditory processes, AP, the Cognitive Dialogue integrates perception, action, and cognition.

5. Research directions

We next discuss some of the questions to be addressed that we believe are important to advance on scene understanding.

5.1. Tools for the late integration of vision, knowledge and language

Consider the problem of activity description, as in the sentence:

A woman cuts the potato with a knife on the table.

Language tools should provide us with two kinds of information. *First*, we need information about the possible quantities in a certain context. In the example above, assuming we know that we are processing a kitchen scene, language should provide the possible objects and verbs used in this setting. Current Computer Vision applications deal with predefined data sets. However, knowledge databases (such as [2]) can provide this information. The Praxicon [7,10], is such a resource, which contains knowledge of common sense everyday activities. This lexical database, obtained by re-engineering WordNet, provides pragmatic relations, on the relations of verbs and objects and it also provides algorithms, which we can use to query domain-specific knowledge, for example if we want to obtain the quantities involved in cooking in a family kitchen.

Second, we need language tools that provide us with contextual relations of different quantities, such as that 'knives' are possibly used for 'cutting', and such activity is often performed 'on tables.' Classical linguistics can build domain knowledge of this kind, and provide information on whether certain combinations are plausible or not. Statistical language tools accessing large text corpora can provide statistics on the co-occurrence of the different quantities in certain domains. We can then use this statistical language information together with statistical information from the visual recognition with classifiers to optimize for scene interpretation. In our own work we have used the statistical language approach for action interpretation in video, where we obtained the probabilities about the co-occurrence of quantities from a large text corpus to generate a sentence description [12]. We also demonstrated the lexical database approach for a robot observing actively humans performing actions to create descriptions that will allow the robot to execute similar actions [11,14].

Interesting questions arise when we realize the dialogue for active agents and construct the models dynamically. As the dialogue proceeds, additional knowledge introduced into the language space changes the expectation for other concepts. Similarly, knowledge creates *expectations in the visual space*, and thus constrains the search space for object and action recognition. For example, if during the dialogue there is a high probability that the human is performing a cutting action, the vision module will not need to run every object classifier to identify the tool used for cutting, but should inspect whether the tool used is one of a small set of cutting tools. Going even further, instead of applying object classifiers, it could instead apply procedures that check the appearance of cutting tools.

5.2. Controlling the dialogue

An essential component of the vision knowledge/language dialogue is that it should guide the attention to expected objects, their locations and attributes and to actions in the scene. We thus need models for the attention mechanism which will predict the order of fixations, i.e. to what and where we should allocate the computational resources. As an example, in [15] we proposed a way to control the Cognitive Dialogue using information theory. At any time, the system has a goal. The goal can be as simple as recognizing a scene by the objects in it, or more complicated as in the recognition of an activity, which is described by many quantities. At each time t, the system must utilize what it already knows in order to pick the optimal question to ask at step t + 1.

We formulated this using Bayesian estimation. The probabilities involved come from the accuracy of visual detectors, and the importance of the individual quantities in describing the scene or activity derived from language. At step t + 1 the criterion for selecting the appropriate quantity is to maximize the information gained about the scene/activity recognition due the response of the added quantity detector, and this can be modeled by the KL divergence between the probability distributions of detecting the activity on the basis of the quantity detectors at step t and the probability distribution of detecting with an additional quantity detector. Adding criteria for how to start the process, such as for example always attending first to moving humans and criteria for finishing the process, we obtain a systematic way of carrying on the dialogue.

5.3. Early integration of vision and knowledge for recognition

At the level of recognition, the integration of vision and language is much more difficult, because written texts usually do not describe the visual appearance of objects and actions. Language can provide some information about attributes: about their color, texture and shape, and this knowledge can be used directly in recognition and segmentation. For example, it is easier to segment the long red object than to generally perform segmentation of the scene. A fertile ground for recognition, we believe, will come from using machine learning techniques that combine visual information in images with information from language, such as attribute information, ontological knowledge, and the affordances (or functionality) of objects.

Actions are compositional in nature. Starting from simple actions occurring on a part of the body, we can compose actions from several limbs to create more complex actions, and we can further combine a sequence of simple actions into activities. Language can be used to enhance the action recognition at the higher levels and its composition from lower-levels onwards. Language can be used to enforce temporal and logical constraints on how actions can be chained together, using a grammar of action [9] that binds sensorimotor symbols (hands, arms, body parts, tools, objects) with language. In this case, the LP will work across all levels, from bi-grams of actions to inferring the most likely activity from the sequence of such bi-grams, using large corpora.

6. Conclusions

One way is to consider vision in isolation as a mechanistic system that learns to detect what is where. This is how vision is studied today for the most part. Another way is to consider vision as part of an intelligent system that reasons (and acts) and can ask questions beyond what and where, such as why, how, who, and many other questions [4,13]. This second way introduces more interesting questions and points to a new theory for the integration of intelligent systems with perception. The Cognitive Dialogue and the tools surrounding it represent our effort toward that direction.

Acknowledgment

The support of the European Union under the Cognitive Systems program (project POETICON++, ICT-288382) and the National Science Foundation under an INSPIRE grant in the Science of Learning Activities program, (grant SMA-1248056) is gratefully acknowledged.

References

[1] D.A. Cruse, Lexical Semantics, University Press, Cambridge, England, 1986.
[2] CYC, http://www.cyc.com/ 2014.
[3] A. Farhadi, S.M.M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D.A. Forsyth, Every picture tells a story: generating sentences from images, Proc. European Conference on Computer Vision (ECCV), 2010, pp. 15–29.
[4] A. Fire, S.-C. Zhu, Using causal induction in humans to learn and infer causality from video, The Annual Meeting of the Cognitive Science Society (CogSci) 2013.
[5] D. Forsyth, T. Berg, C. Alm, A. Farhadi, J. Hockenmaier, N. Loeff, G. Wang, Words and Pictures: Categories, Modifiers, Depiction and Iconography, Cambridge University Press, 2009.
[6] A. Gupta, A. Kembhavi, L.S. Davis, Observing human–object interactions: using spatial and functional compatibility for recognition, IEEE Trans. PAMI 31 (10) (2009) 1775–1789.
[7] Poeticon, http://www.poeticon.eu 2012.
[8] G.A. Miller, C. Fellbaum, WordNet now and then, Lang. Resour. Eval. 41 (2) (2007) 209–214.
[9] K. Pastra, Y. Aloimonos, The minimalist grammar of action, Philos. Trans. R. Soc. B Biol. Sci. 367 (2011) 103–117.
[10] K. Pastra, E. Balta, P. Dimitrakis, G. Karakatsiotis, Embodied language processing: a new generation of language technology, Proc. AAAI Int. Workshop on "Language–Action Tools for Cognitive Artificial Agents" 2011.
[11] D. Summers-Stay, C.L. Teo, Y. Yang, C. Fermüller, Y. Aloimonos, Using a minimal action grammar for activity understanding in the real world, RSJ International Conference on Intelligent Robots and Systems (IROS), 2012, pp. 4104–4111.
[12] C. Teo, Y. Yang, H. Daume, C. Fermüller, Y. Aloimonos, Towards a watson that sees: language-guided action recognition for robots, IEEE International Conference on Robotics and Automation (ICRA), 2012, pp. 374–381.
[13] P.F.M.J. Verschure, Distributed adaptive control: a theory of the mind, brain, body nexus, Biol. Inspired Cogn. Arch. 1 (2012) 55–72.
[14] Y. Yang, A. Guha, C. Fermüller, Y. Aloimonos, A cognitive system for understanding human manipulation actions, Adv. Cogn. Syst. 3 (2014) 67–86.
[15] X. Yu, C. Fermüller, C. Teo, Y. Yang, Y. Aloimonos, Active scene recognition with vision and language, Proc. International Conference on Computer Vision 2011.