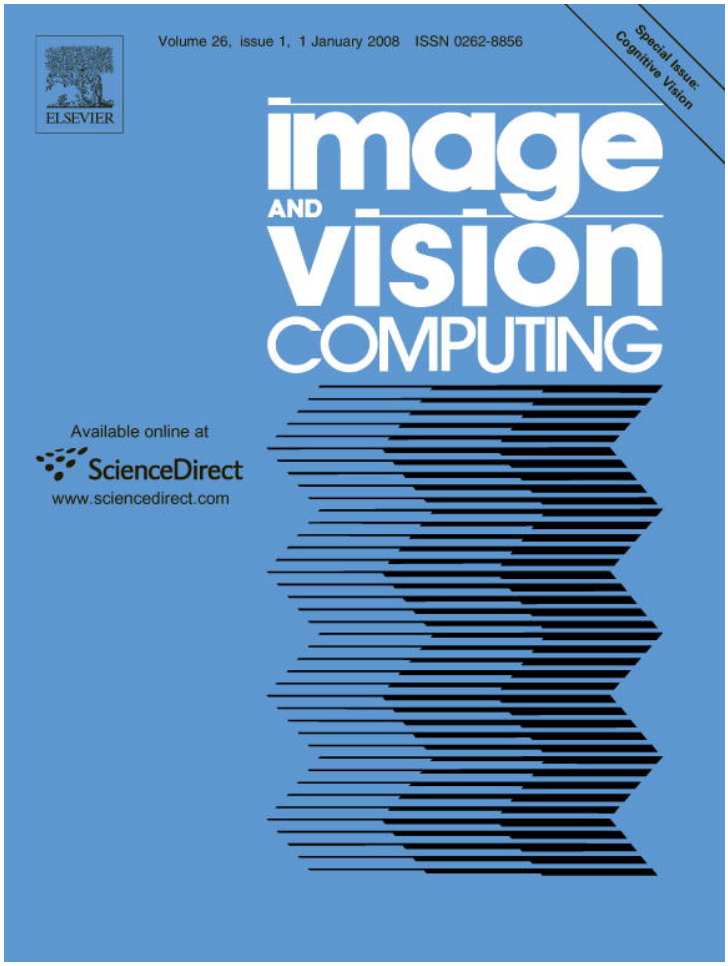

Available online at www.sciencedirect.com
ScienceDirect
Image and Vision Computing 26 (2008) 1–4
www.elsevier.com/locate/imavis

Editorial

Image and vision computing special issue on cognitive vision

Special issues of scientific journals reflect critical points in the evolution of a discipline. They are snapshots taken at a particularly important moment when several of the threads that make up a discipline are woven together to reveal a new pattern, one that will reoccur repeatedly in the future as the discipline matures. This special issue of *Image and Vision Computing* is dedicated to cognitive computer vision. The importance and timeliness of taking this snapshot was first noticed by Hilary Buxton, Professor of Visual Intelligence at the University of Sussex, and she was the driving force behind all the work that it took to bring together the papers collected in this special issue. Many things inspired her to do this but I believe that two important occurrences were pivotal. The first was a lively seminar on Cognitive Vision Systems organized by Henrik Christensen and Hans-Hellmut Nagel held at Schloss Dagstuhl in October 2003 (Seminar 03441). The second was the far-sighted decision by the European Commission in Luxembourg to launch a funding programme dedicated to this topic, a decision that fostered the emergence of a completely new thread of investigation in the computer vision community. It is no accident that the research described in 7 of the 10 papers in this special issue can trace their roots to projects funded by this programme.

What exactly do we mean by cognitive computer vision? Definitions, especially concerning matters that have not yet fully evolved, can be dangerous because they seek to place boundaries when we do not yet know the full extent of the issues at stake. So, it is best perhaps not to attempt one here and focus instead on the issues that concern cognitive vision in that hope that it will provide a rough context in which to place the contributions contained in this issue.

Cognition implies an ability to understand how the world around us might possibly be, both now and at some future time, and to take this into consideration when determining how to act. There are two aspects to this: anticipating or predicting future events and being able to interpret a visual scene without having complete data. Aaron Bobick put it succinctly at a cognitive computer vision summer school in 2005 when he said that 'Cognitive vision is a lot about being able to assert that something is there, given very little visual evidence, and even perhaps despite evidence to the contrary'. To be able to achieve this, a cognitive system should have the capacity to acquire new knowledge and to use it to fill in the gaps that are present in what is being made immediately available by the visual sensors: to extrapolate in time and space to achieve a more robust and effective understanding of the underlying behaviour of the sensed world. In the process, the system learns, anticipates, adapts, and improves. These two characteristics of adaptability and anticipation are the hallmarks of cognition, in general, and cognitive vision, in particular.

In this issue, the paper by Bauckage et al. – *The visual active memory perspective on integrated recognition systems* – presents the visual active memory (VAM) approach to building an interactive cognitive system that can operate with humans in everyday surroundings. The VAM approach deals directly with the problems that have plagued traditional computer vision systems, *viz*: the impossibility of designing *a priori* a complete model of all possible objects, activities, and actions, and the need to build systems that can deal dynamically with previously unseen image data, characterize previously unseen objects, and address *ad hoc* tasks, not just ones that have been pre-specified. The VAM approach exemplifies several aspects of cognitive vision: it adopts an integrated systems approach; it is adaptive and incrementally acquires visual models, generating knowledge by tightly coupled cooperation between processes of perception, reasoning, and learning; it exploits prior models to get the system started; it is interactive and communicates when necessary with humans; it is founded on the principle that the system must be able to 'play around' with its own state and thereby be self-modifying; and it facilitates multiple computations through distributed processing and easy integration of new modules. The incremental acquisition of visual models is achieved by coupling model acquisition and recognition so that object models are acquired over a period of time and are modified and improved on the basis of recognition performance. The system is structured as an XML database with four hierarchically organized memory systems (pictorial, fea-

0262-8856/$ - see front matter
doi:10.1016/j.imavis.2007.09.003

ture-based, episodic, and categorical). Recognition of previously seen objects, activities, and contexts can trigger a re-encoding of learned models. The VAM approach exploits a human in the learning loop, both to provide assistance with labelling, and to alter the physical configuration of the scene and cameras.

The paper by Liebe, Ettlin, and Schiele – *Learning semantic object parts for object categorization* – also focusses on adaptive knowledge acquisition, relaxing the requirement for visually relevant information to be specified by the designer of the cognitive vision system. Instead of having to specify *a priori* the parts that comprise objects for use with an appearance- and model-based object categorization process, this paper shows how a cognitive vision system can learn intermediate representations of object parts through progressive application of grouping processes. Beginning with a conventional similarity-based class-specific appearance features (i.e. visually similar image patches), the system then clusters these to produce sub-part representations on the basis of inter-image 'co-location' of their features (consistent placement relative to the associated object), finally grouping these sub-parts on the basis of intra-image 'co-activation' (consistent activation or mutual presence in a given spatial neighbourhood). This process produces semantically relevant parts many of which would be identified as a part by a human observer, but others (e.g. consistent shadows) which might not. The level one visual appearance features are used in a Generalized Hough Transform to hypothesize the presence of a given part (this can be accomplished because each patch feature has associated with it its location with respect to the object centre). This hypothesis is verified using a Bayesian Network to combine top-down structural knowledge with the automatically learned sub-parts and parts that results from the tiered grouping process. This process requires some human intervention during the training phase: specifically the identification of the location of the patch feature with respect to the object centre.

The paper by Skočaj and Leonardis – *Incremental and robust learning of subspace representations* – continues on the theme on learning new knowledge. Since cognitive vision systems do not assume all the knowledge required to carry out their tasks, they need to be able to learn and, more specifically, the need to be able to learn in an incremental, continuous, open-ended, and robust manner, with the processes of learning and recognition being interleaved, and with both processes improving over time. To enable this, the representations which model the observed world need to allow the efficient assimilation of new information without requiring access to previously observed data, while preserving previously acquired knowledge. Since the updating process (i.e. the learning) will normally be effected autonomously without supervision, e.g. when the system itself recognizes something, the learning technique needs to be able to distinguish between good and bad data, otherwise bad data will corrupt the representation and cause errors to become embedded and to propagate. In other words, learning needs to be robust and should only use data that is consistent with what has already been learned. The authors deal in this paper with appearance-based representations. These approaches avoid the problem of decoupling object shape, reflectance properties, object pose, and scene illumination by systematically observing the object to be learnt under different viewing and illumination conditions. Principal component analysis (PCA) provides a compact encoding of an otherwise very large representation space. However, standard PCA is not suitable for incremental robust learning: it is typically used in batch mode with all training images being provided in advance. Consequently, there is no way to weight either individual image or individual pixels to reflect their relevance to the object being modelled or their consistency with what was previously learned. This paper introduces an extension of conventional PCA learning which incrementally updates the principal subspace, which can temporally weight individual images and spatially weight individual pixels, and which can determine the consistency of input (training) data to ensure it is compatible with previously acquired knowledge. This incremental approach has the significant advantage that images can be discarded once they have been used in the learning process. The approach, therefore, is very well matched with the exploratory and development needs of cognitive vision. Since the technique determines the consistency of later training data, it is important that the image data used to build the initial representation should be reliable and should be heterogeneous, i.e. the data should reflect a good variety of object viewing situations.

The predictive element of cognition, and especially the expectation-driven guidance of perception, is addressed in the paper by Sage, Howell, Buxton, and Argyros: *Learning temporal structure for task-based control.* This is done in the context of an adaptive skin colour visual hand-tracker and a variable length Markov model (VLMM) functional gesture recognition system. The authors' principle point is that a cognitive vision system should be able to learn a behavioural model (in this case the functional gestures someone makes when picking up something or reaching to retrieve an object). This model can then be used to assist the tracking module when it has insufficient data to operate e.g. such as is the case when the hand is occluded for long periods mid-gesture. This provides a form of learned expectation-driven predictive control. To accomplish this, the authors present an extension of a variable length Markov model (VLMM) which can model continuous input data (hand coordinates) to learn structural motion, i.e. functional gestures, and which exploits the stochastic generative properties of the VLMM to extrapolate the missing positional data when the tracker is unable to provide it.

Arens, Gerber, and Nagel also highlight the key idea of the use of knowledge to reason about the potential (future) behaviours of agents in their paper *Conceptual representations between video signals and natural language descriptions* in which they describe the evolution of a cognitive vision

system which can produce textual descriptions of inner-city traffic behaviour from video data. This behaviour is achieved through a succession of processes, grouped in three layers: the quantitative (numerical) layer, conceptual (symbolic) layer, and natural language layer. In the conceptual layer, behavioural knowledge is represented using situation graph trees (SGTs) which describe the situation an observed agent is in, any associated actions, and possible future situations. Thus, SGTs represent not just the current observed state but also potential behaviours of agents. The most recent generation of the system allows the SGTs to be created and modified graphically by the designer. The processes in the three layers all make use of background knowledge. A characteristic feature of this system is that this background knowledge is provided explicitly by the system designer. The authors argue that this reliance on engineered knowledge bases allows one to minimize unexpected problems and work on refining the processing architecture; one can later revert to learning background knowledge once the system components have gained sufficient robustness.

In *Enhanced tracking and recognition of moving objects by reasoning about spatio-temporal continuity*, Bennett, Magee, Cohn, and Hogg introduce a framework that exploits logical reasoning with explicit symbolic qualitative spatio-temporal continuity constraints to significantly improve the performance of a typical imperfect blob-based visual tracking and colour histogram-based recognition. The proposed method is independent of many of the problems that cause trackers difficulty when confronted with occlusion: how accurately the object dynamics and appearance are modelled, the time-scale over which the occlusion occurs, the complexity of the object behaviour during occlusion, and the degree of occlusion. Rather than trying to disam-biguate the occluding and occluded objects with advanced image processing, their strategy is instead to reason about the occlusion taking place, explicitly increasing the tracker ambiguity where errors are likely (e.g. at the onset of and during occlusion) by merging overlapping blobs, and then reasoning about spatio-temporal continuity to resolve the ambiguity. After training the classification algorithm to produce exemplar colour histograms for each hand-labelled object in a training video sequence, these exemplar histograms are used to assign probabilities of each object to the bounding box of all tracked blobs. This enables the construction of first a directed graph representing the evolution of the blob bounding boxes (some of which will correspond to a single object, others to multiple objects) and subsequently a higher-level set of spatio-temporal envelope. These envelopes capture the evolution of the bounding boxes in a way that allows the application of a Prolog error reasoning and consistency checking module to identify the solution that maximizes the statistical correlation of the tracker and classification modules. It also identifies the most globally consistent blob behaviour with respect to two spatio-temporal constraints: exclusivity (an object cannot be in more than one place at the same time) and continuity (an object's motion must be continuous).

The paper by Neumann and Möller – *On scene interpretation with description logics* – outlines a conceptual framework for knowledge-based scene interpretation, discussing how it can be realized with description logics augmented by Bayesian Networks to effect selection between several plausible incrementally generated models. Description logics (which are based on a subset of first order predicate calculus) are used to represent and reason about high-level concepts such as spatio-temporal object configurations and events, capturing common-sense knowledge, past viewer experience, and non-observable contextual information. Their goal is to build a system that can produce a consistent interpretation of the visual scene by incrementally constructing partial models which may explain the visual information present in a possibly partially complete geometrical scene description. Since the space of consistent interpretations can be very large and since the knowledge-based framework is not suitable for expressing preferences of one interpretation over another, the authors outline a probabilistic framework which combines Bayesian Networks and description logics to guide interpretation and help select between multiple possible models. Thus, the preference measures could possibly be learnt rather than having to be hand-crafted as rules.

The system described in the paper by Maillot and Thonnat – *Ontology-based complex object recognition* – illustrates the combination of explicit knowledge bases created by a human designer and machine learning in a cognitive vision system designed to categorize complex isolated objects. The knowledge base makes explicit symbolic domain knowledge while these symbols are grounded through learning by associating these high-level symbols with image features. Once a domain taxonomy – a specification of domain classes and their specialization and part-whole relations – has been developed by interacting with an application domain expert, a visual description of these domain object classes is developed using a pre-defined visual concept ontology based on texture, colour, and spatial concepts by training a set of classifiers to associate or recognize each concepts with features derived from prototypical manually segmented and annotated image samples The result is a domain knowledge base. Finally, object categorization is effected by segmentation, feature extraction, and recognition using the visual concept classifiers trained on the sample data.

In their paper *Attention links sensing to recognition*, Rothenstein and Tsotsos observe that attention has been neglected to a significant extent by the vision community, often because vision systems make assumptions that 'reduce or eliminate the need for attention'. These include targetted camera set-ups, clean backgrounds to facilitate simplified segmentation, assumptions about image features, prior knowledge of the task domain, scene content, and object behaviour. The authors note that these assumptions do not apply in everyday circumstances, i.e. exactly

where cognitive vision is needed. Since cognitive vision is concerned with building systems that limit the amount of *a priori* knowledge required for robust performance, attention is pivotal to cognitive vision. Their paper presents a comprehensive survey of approaches to computational attention and it discusses the Selective Tuning (ST) model in particular. Since ST uses a winner-take-all strategy selection mechanism, it is particularly consistent with cognitive architectures that deploy multiple concurrent perception processes which must be modulated to produce the appropriate cognitive behaviour. At present, ST and the other attention models surveyed provide restricted selection capabilities (e.g. regions of the visual field, features of interest, and the parameters of low-level operations). They cannot (yet) select higher-level information such as relevant task-specific objects, events, or system behaviours from a knowledge base.

There is considerable debate in the cognitive systems community about whether or not cognitive systems need to be embodied and, indeed, what it means to be embodied in the first place. My own paper in this issue – *Cognitive vision: the case for embodied perception* begins with an attempt to set the scene for the core argument by briefly reviewing the several paradigms of cognition, including cognitivist, emergent, and hybrid systems, and their respective strengths and weaknesses. The central issue of the paper is whether or not embodiment is a necessary requirement for cognitive vision systems and, if so, what that means in practice. The paper looks at the case for embodiment from both the cognitivist side (there being no requirement for embodiment) and the emergent side (embodiment being essential). The paper then argues a case for embodiment independently of the paradigm one adopts, an argument that hinges upon the two assumptions that a visually perceptive cognitive system is autonomous and that it acquires empirical knowledge of its environment. On the basis that at least one of the two arguments put forward for the necessity of embodiment is valid, the paper then proceeds to look at what embodiment actually entails and what types of embodiment may be usefully considered in the context of a cognitive vision system. This leads the paper to consider the implications of accepting the necessity for embodiment of whatever type: the balance between phylogeny and ontogeny – i.e. between innate and developed capabilities and knowledge – the configuration of a cognitive vision system and the limitations on the rate of learning and development imposed by real-time system environment coupling and embodied development. Before concluding and summarizing its arguments and findings, the paper then moves on to consider what we can learn from the study of natural cognitive systems to help us answer the many questions that have been posed.

David Vernon
Etisalat University College, United Arab Emirates
E-mail address: david@vernon.eu