A Layered Active Memory Architecture for Cognitive Vision Systems

Ilias Kolonias, William Christmas, and Josef Kittler

Centre for Vision, Speech and Signal Processing,
University of Surrey,
Guildford GU2 7XH, UK
{i.kolonias,w.christmas,j.kittler}@surrey.ac.uk
WWW home page: http://www.ee.surrey.ac.uk/CVSSP/

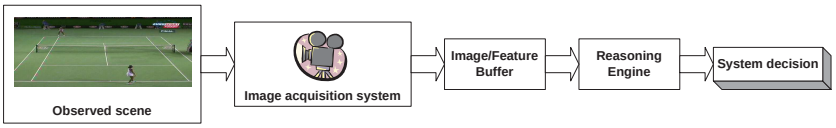
Abstract. Recognising actions and objects from video material has attracted growing research attention and given rise to important applications. However, injecting cognitive capabilities into computer vision systems requires an architecture more elaborate than the traditional signal processing paradigm for information processing. Inspired by biological cognitive systems, we present a memory architecture enabling cognitive processes (such as selecting the processes required for scene understanding, layered storage of data for context discovery, and forgetting redundant data) to take place within a computer vision system. This architecture has been tested by automatically inferring the score of a tennis match, and experimental results show a significant improvement in the overall vision system performance — demonstrating that managing visual data in a manner more akin to that of the human brain is a key factor in improving the efficiency of computer vision systems.

1 Introduction

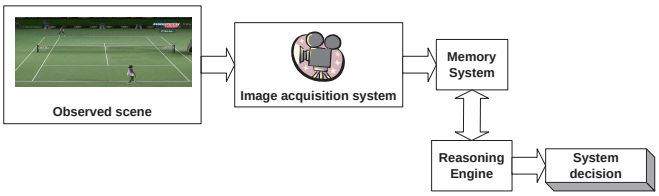
Visual perception is an area of computer vision witnessing considerable progress during the last few years. Novel learning algorithms (such as Support Vector Machines, AdaBoost and Multiple Classifier Systems) and models of geometric invariance, picture formation and noise have enhanced the capability of machine perception systems. Nevertheless, visual perception systems are built using the signal processing paradigm for information flow: visual data are captured; low-level feature extraction algorithms detect points or areas of interest in the images; and a decision mechanism determines whether a predefined pattern exists. Since computer vision applications tend to be developed as application-specific solutions, without serious consideration about generic data pattern storage or fusing information from other sources, a buffer accessible at each level provides data storage — while the information flow only allows forward interaction between building blocks. This information processing strategy is illustrated in Figure 1(a).

While very simple, this architecture has been employed in a variety of visual tasks [1–3]. It is clear, however, that backward interaction between the various levels of data processing is not supported; therefore, it is incapable of exploiting contextual information for reasoning about a scene. Moreover, the lack of a generic data management scheme causes serious problems regarding the storage and (especially) the fusion of information from different sources and/or levels of abstraction. This precludes the design of modular computer vision systems where the system itself can decide, in a unified and elegant manner, the data processing

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de



(a) Standard computer vision system



(b) Cognitive vision system

Fig. 1. Information processing strategies for visual perception systems.

strategy required for the given task. As exploiting context for reasoning within a known environment is crucial for injecting cognition into a visual perception system, a memory architecture enabling the computer vision system's reasoning engine to discover contextual links within the data needs to be developed. To this end, the memory system must allow re-examination of previous data in the light of newer evidence — while a contextual reasoning framework will ensure that a priori knowledge about the scene is also taken into account. This tightly coupled memory–contextual reasoning framework paradigm (as opposed to the signal processing one) supporting the introduction of cognitive capabilities in computer vision systems is depicted in Figure 1(b).

In designing a memory system for use with visual perception systems, other issues must be addressed as well — one being the amount of memory available for visual data processing. A memory architecture not discarding low-level visual data is seriously limited with regard to its applicability in real-world applications, as it will only be able to hold video data for a limited amount of time. Thus, discarding data without compromising the system performance is essential. An active memory approach can solve this problem; memory resources can be freed as soon as the system decides there is redundant data. However, deciding what is redundant in a cognitive vision system can be tricky, since, for visual data to be deemed as such, it has to bear no importance to the current state of the scene. This is decided by the system's contextual reasoning engine — which demonstrates that the efficient handling of contextual information is the most distinguishing feature between biological cognitive vision systems and their machine-based counterparts. Consequently, developing a memory infrastructure inspired by biological cognitive systems for handling both sensory data and corresponding high-level abstractions is a natural choice.

2 Data fusion in biological and computer vision systems

2.1 Biological insights on information fusion and cognition

A large body of work in physiology has been devoted to understanding the mechanisms by which cognition is achieved in living organisms. Even since the

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

late 1970s, the concept of *'the unity of senses'* [4] has been studied; this suggests that stimuli are usually perceived by more than one sense, allowing humans to perceive the same thing in different ways. Biologists concluded that our senses share common perception mechanisms, and that a common representation of input stimuli is adopted to integrate and interpret multi-sensory information via a single, common perception-enabling mechanism.

Neurologists have independently provided neuron connectivity models for sensory data fusion supporting this theory [5], revealing that no interaction between signals transmitted from the senses to the superior colliculus (the part of the brain where sensory data arrives first) occurs — but neurons leaving the superior colliculus are multi-sensory. The superior colliculus also receives information from the cerebral cortex — the part of the brain that modulates behaviour. Therefore, information fusion using a *perception-action coupling* paradigm to take *contextual information* into account takes place in the superior colliculus. This provides flexibility in the information fusion scheme used in each case *and* enables the detection of conceptually important events even by a set of weak cues.

Moreover, MEG and EEG scans studied in [6, 7] reveal the existence of a short-term episodic memory for encoding visual stimuli. These findings indicate that, for recognising known visual patterns within perceived visual information, humans do not use raw images; instead, features from the input visual data are extracted and matched against the known pattern. Clearly, the human brain being a highly parallelised information processing system, it has an important advantage over state-of-the-art computer vision systems. However, it is the information processing paradigm adopted in biological vision systems that allows them to achieve a level of efficiency so far beyond that even state-of-the-art machine vision systems can reach — not necessarily the low-level visual feature extraction techniques employed.

2.2 Cognitive and behavioural models on sensor data management

Biological studies provide interesting insights into how the human brain achieves sensor fusion. Cognitive and behavioural sciences, however, investigate why sensor fusion is required for perception. *Perceptual modes* [8] are an important notion in understanding cognitive mechanisms — suggesting that, when the intention of the agent changes, so does the interpretation of a given stimulus. For example, a person notices different things when entering a room just to see what is in it, as opposed to what he/she will notice when searching for something. This indicates the presence of an adaptive fusion mechanism for sensory data — proving that perception-action coupling is present in living organisms.

Sensor fusion must efficiently handle discordances among inputs. This can be achieved in one of the following ways [9]: re-calibrate the sensors until the perceptual goal is met *and* sensory input is consistent; suppress offending sensory data; or avoid attaching any spatiotemporal correspondence across sensory data. Therefore, a closed-loop control topology for sensor fusion has to be developed; bottom-up or top-down approaches for associating concepts to sensory input may be useful in understanding perceptual processes, but the input-percept hierarchy

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

in real-world scenarios is not always as straightforward as that, especially when feedback is involved.

The core behaviour of biological cognitive systems can be divided in two types: an *investigatory* mode, where the system looks for perceptual information relevant to a given cognition task, and a *performatory* mode, for performing the task [10]. Computer-based cognitive systems simulate these two modes by *bootstrapping* and *normal operation* (deployment) respectively. Applying this strategy in a computer-based cognitive system prevents the (computationally expensive) bootstrapping process from being constantly invoked; yet the normal operation process can re-use it (in form of an intelligent agent) for re-adjusting to sensory input.

Cognitive psychologists have experimentally demonstrated the presence of a short-term working memory in biological cognitive systems [11] — in contrast to the long-term memory used for learning conceptual entities. However, it is possible to utilise this buffer for visual processing as well as representation [12]. Data in the short-term working memory are stored very briefly, but may be instantly recalled in full detail (with all their associated attributes). Storage in the long-term memory, however, is associative, relating large numbers of differing items based on their co-occurrences rather than their inherent attributes. Recall from the long-term memory is thus slow (lacking immediate access to the attribute data) and completely relies on providing sufficient retrieval cues related by association to the item under consideration. The discrete patterns represented in long-term memory are high-level abstractions of sensory representations within the short-term working memory, and are originally allocated on the basis of how often that particular set of attributes has occurred within the working memory [13]. There is hence an inverted relationship between memory retention and interpretative level amongst human subjects.

2.3 Data fusion, management and reasoning in computer vision

The findings described above indicate that the memory of a biological cognitive system is two-layered; there is a 'working buffer' (where low-level data is retained in detail for a limited period of time) and a long-term memory (where conceptual processing results are stored for as long as the system sees fit). Whereas the structure and function of memory in cognitive applications is crucial for successfully deploying the overall system, it has not been as thoroughly investigated by the computer vision community as, for example, feature extraction and object/action recognition.

A model for sensor fusion for different levels of information (raw data, features, or decisions about input content) is described in [14]. As real-world sensors cause information *fission* due to physical constraints, a suitably designed *fusion* process must counteract this. Ways of fusing different types of visual processes to enhance the robustness of active vision systems are studied in [15]. Still, the notion of memory as a storage buffer is *only superficially* covered, the focus being on information fusion for improving the decision-making process.

In [16], an *active memory* serves as a basis for fusing information across modalities and facilitating reasoning on perceived data for deploying recognition

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

systems. This model incorporates the basic apparatus enabling both *intrinsic* (tightly linked) and *extrinsic* (loosely linked) processes to manage stored data. *Forgetting* is cited as an example of an intrinsic memory process, while *consistency validation* (i.e. reasoning) is a typical case of an extrinsic process, thus suggesting that cognitive vision tasks can be seen as interactive memory processes. While this memory architecture is quite flexible, it still operates on a single layer — while a memory architecture for use with a cognitive vision system also requires a low-level visual data buffer to be present.

Nonetheless, this is a step forward from what most state-of-the-art systems apply — which is normally direct fusion and decision-making in the same step, as [14] seems to suggest. A number of decision-making schemes have been employed in computer vision — including Bayesian Networks, Dempster-Shafer theory, Neural Networks, Self-Organising Maps, or particle filters. As researchers have generally opted for a single-layer decision-making mechanism, they have not addressed the possibility of directly applying their systems to different domains, or of efficient storage of visual data and the results of its analysis.

3 A Multi-Layer Memory System for Cognitive Vision

In this section, a novel *memory infrastructure* used for the spatio-temporal processes related to a cognition task where the observed process *is reasonably continuous over time* is presented, and applied in a computer vision task to facilitate the storage of conceptual results and the injection of cognitive capabilities for a scene interpretation and understanding problem — annotating off-the-air tennis match broadcasts. The example application demonstrates the main concepts and building blocks enabling cognition in computer vision systems.

3.1 Logical Architecture

As mentioned earlier, a crucial feature of cognitive vision systems is the presence of a multi-layer, flexible memory architecture — enabling the management of its content to be dependent on the conceptual importance of the content itself, as well as facilitating information fusion for decision making at all levels. Drawing inspiration for the design of a computer memory system from biological systems will be a good starting point for enabling cognition in computer vision. The basic layout of the human cognition mechanism is illustrated in Figure 2(a).

Cognition in humans utilises three levels of memory storage: a sensory information buffer, the short-term, working memory (which is further subdivided in two parts, handling low-level feature data and elementary concepts respectively) and the long-term memory. The sensory information buffer handles data for only a very limited time frame (less than a second) and is used at the lowest level of human perception, allowing the brain to process the input stimulus and extract potentially important low-level features. Extracted features are then stored in a short-term memory repository — from this stage on, human cognitive processes have full control of how feature information will be managed. In a typical recognition scenario, the data stored here is typically available for *only a few seconds*, allowing basic object/action recognition and visual attention tasks to take place. However, the latter requires *feedback* from the cognitive centre to the

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

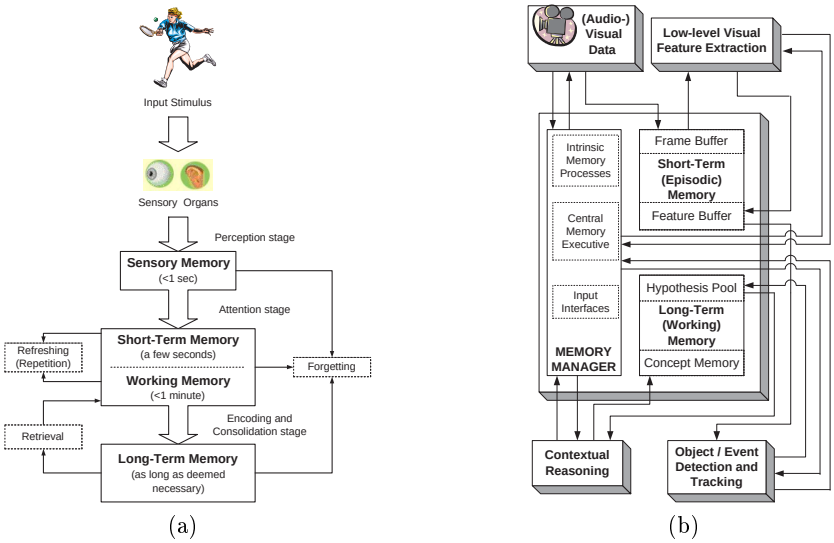


Fig. 2. Functional comparison between (a) the memory architecture of the human brain, and (b) the proposed system.

sensors — which is essential to the adaptability of cognitive vision systems to different environments.

For handling more abstract concepts, a higher level of memory also exists. In this case, memory contents are preserved over a somewhat longer term (up to a minute), while the requirement for flexibility in representing and fusing memory content becomes apparent; memory contents (and their relationships) are much more application-specific. At this level, memory contents are treated as *hypotheses* about the scene evolution, and combined with other hypotheses (from different sources or time scales) to assess whether they represent the scene content. Hypotheses plausible for the given scene are then stored in the long-term system memory. At this level of abstraction, the memory repository is clearly application-specific, its content being a tightly-structured set of concepts concerning entities and interactions in the observed scene. This is the highest level of compactness achievable for describing the observations made, and the data is retained for as long as the system requires it. Reaching such levels of compactness and abstraction in visual data description are the most important benefits of injecting cognition in vision systems, allowing operations such as intelligent data querying and re-enacting the scene evolution from a minimal description data set.

Figure 2(b) shows the conceptual architecture of the proposed memory infrastructure. Two levels of memory storage exist — a short-term and a long-term component. The former operates equivalently to the human short-term sensory and low-level working episodic memory, whereas the latter functions similarly to the working and long-term human memory. Both short-term and long-term components are further divided, each into two parts. The short-term memory consists of a *frame buffer* and a *feature buffer*. The frame buffer contains raw image data and retains them for a very limited amount of time. The feature

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

buffer is used by low-level visual feature extraction algorithms to store their results for elementary object/action recognition tasks. The long-term memory component is divided into a *hypothesis pool* and a *concept memory*. The hypothesis pool accumulates plausible hypotheses about high-level entities or concepts; therefore, it operates like the working memory of a biological cognitive system. Finally, the concept memory retains high-level concepts verified by combining elementary hypotheses and applying the appropriate contextual constraints. At this level of abstraction, concepts form a tight description of the perceived scene and are treated as factual data, suitable for future reference.

The inherent hierarchy in cognitive tasks also necessitates the presence of a layered structure for information exchange between different categorical domains. To this end, a top-level XML file, outlining the tasks undertaken by individual modules (and thereby annotating interactions between them) is supplied; each module is registered in this collection, and all modules' input data sources are described. This convention allows the system to decide on its own whether a module needs to be executed, thus saving computational resources if low-level visual features are available. Finally, this architecture allows the system to be easily re-configured for different cognition tasks, provided the modules required for the new task are present. Modules interacting with the memory also need to provide information about the semantic level of content they output, which correspond to the layers present in the proposed architecture.

3.2 Memory Content Organisation

Within the memory system, input data must be suitably represented to allow for the implementation of reasoning capabilities by external processes. As arbitrary input data structures must be efficiently handled, memory data are stored as XML documents. Each process interacting with the memory provides its own XML schema for the data it produces and stores within the memory, so that other processes can access that data as well.

The size of data items may sometimes pose practical issues. Hence, when large items (most notably, images) are to be stored, a slightly different strategy is followed; the actual data are *stored separately* in files and only *references* to these files are inserted into the XML memory files instead. This makes the complete memory system resemble a repository, in which feature and concept data (which are smaller in size and their structure can be described via XML schemas), are stored inside the XML-based memory, while large data chunks are stored in separate files. Linking those additional memory resources to the core XML-based memory is based on the time instant the data is produced, or the duration of time for which the data is relevant. This storage convention allows us to handle large data items within the memory system more easily; the data itself is subject to the same memory processes as the data stored within the XML memory documents.

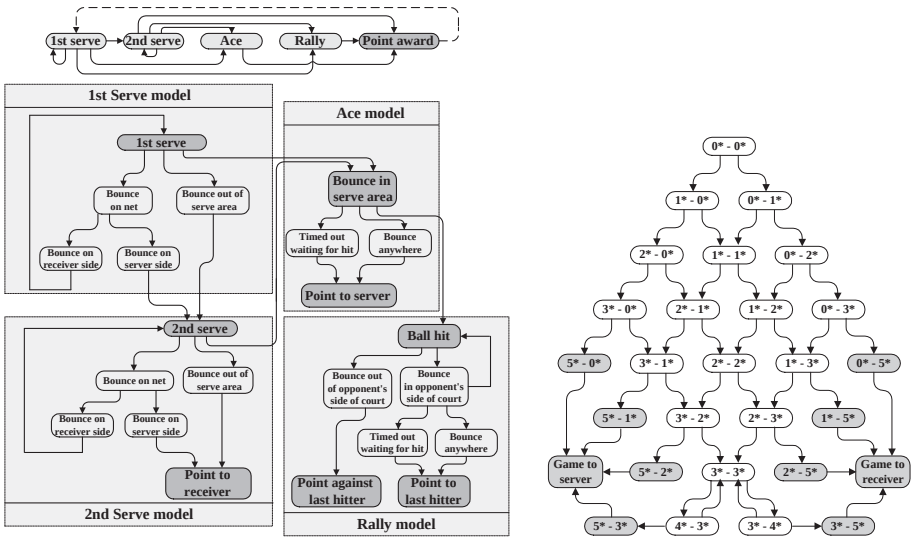
The memory data are structured using *observation Directed Acyclic Graphs* (oDAG's), where each oDAG refers to a single categorical domain. This choice was made due to the fact that the temporal link is prevalent in cognitive vision tasks, as the evolution of the scene itself is, as a concept, synonymous with

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

the discovery of relations and interactions among its entities in the temporal domain. Another reason is the ease with which the data within an oDAG can be manipulated for implementing reasoning: input observations are accessed (and modified) by traversing the graph; adding data at some point is done by adding sub-graphs at that point; and pruning the graph at a point removes the data stored at that point. Finally, observation chains can be easily manipulated by reasoning tools (such as Hidden Markov Models) for learning underlying concepts from data. In this work, a unified Bayesian framework for contextual reasoning at any semantic level [17] has been deployed.

4 Evolution Tracking of Tennis Videos as a Cognitive Process

Enabling cognition and aiding reasoning in context for computer vision systems are the reasons for developing the proposed memory infrastructure. In tennis, the contexutual information conveyed by its rules [18] can assist a visual perception system to decide the evolution of a match, as reflected in its score. The conceptual diagrams reflecting the rules of tennis for awarding points and games are shown in Figures 3(a) and 3(b) — sets and the match are awarded similarly to games.



(a) (b)

Fig. 3. Evolution models in tennis: (a) Point award, (b) Game award from points. The score is noted as <**Server**> — <**Receiver**> and interpreted as follows: **0*** - 0 points ('Love'); **1*** - 15 points; **2*** - 30 points; **3*** - 40 points; **4*** - Advantage; **5*** - Game

Thus, to extract the score in tennis, the previous score line and all play events since the last point was awarded are required. Nonetheless, the extraction of these high-level concepts entails a number of low-level visual feature extraction processes and object tracking/action recognition — for objects like the ball and the players. This information is fused and object interactions are detected — such

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

as the ball being hit, bouncing on the court, or the officials making decisions. Sets of hypotheses about the match evolution are formulated, and the most likely is considered to be the outcome so far. The main tasks required, the sequence in which they are performed, and the corresponding memory levels where they output their results are illustrated in Figure 4.

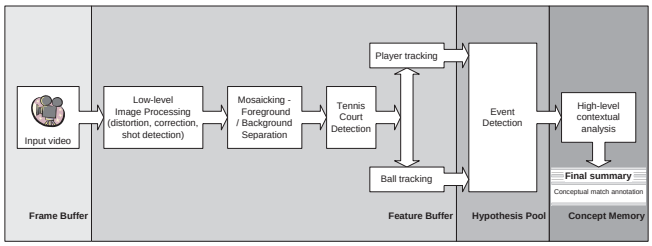


Fig. 4. Sequence of detection and tracking tasks for the tennis annotation system, and the memory levels the corresponding output is stored at. Input comes from all levels of memory storage to the left of, and including, the current level.

The scheme described above has been tested on 40 minutes of broadcast tennis play from the Women's Final of the 2003 Australian Tennis Open, as well as 1 hour of the Men's Final from the same tournament. A total of 80 and 136 'play' shots were processed respectively, each having one of 6 possible outcomes; no play; bad serve by either player; point awarded to either player; or incomplete play. The first set of experiments was done *without* using the feedback capabilities of the proposed memory framework; therefore, *only* the signal processing paradigm for computer vision tasks (allowing *shot-by-shot* video analysis) was feasible. The second set of experiments was carried out with the same parameters for low-level feature extraction as the first one, but *enabling* the memory system's capabilities for information re-assessment and feedback; the results are shown in Table 1.

Test sequence (all from Australia 2003 Tennis Open)	Memory system disabled (shot-by-shot reasoning)			Memory system enabled (point-by-point reasoning)		
	Total	Correct	Rate	Total	Correct	Rate
Women's Final	80	56	70%	48	42	87.5%
Men's Final	136	93	68.38%	99	74	74.75%

Table 1. Summary of system performance and error causes

Thus, simply concatenating the outcome of each shot into an overall description is *not* an adequate method of tracking the evolution of the tennis match, as not all available information is harnessed. However, using the proposed memory infrastructure for discovering and exploiting context has resulted in a significant performance boost for the overall vision system.

5 Conclusions

In this work, a memory infrastructure allowing cognitive processes to take place in computer vision systems has been proposed. Its most distinguishing feature is its ability to manage data in a way conducive to discovering and exploiting

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8

contextual links among them. The combined visual perception/ active memory model is both reliable and readily adaptable to a wide range of cognitive tasks that require analysis at a number of different semantic levels. The proposed system has been evaluated on the analysis of tennis video material, with very encouraging results.

References

1. Petkovic, M., Jonker, W., Zivkovic, Z.: Recognizing strokes in tennis videos using Hidden Markov Models. In: Proceedings of Intl. Conf. on Visualization, Imaging and Image Processing, Marbella, Spain. (Sep 2001)
2. Denman, H., Rea, N., Kokaram, A.: Content-based Analysis for Video from Snooker Broadcasts. ELSEVIER Computer Vision and Image Understanding **92**(2-3) (November-December 2003) 176–195
3. Assfalg, J., Bertini, M., Colombo, C., Bimbo, A.D.: Semantic Annotation of Sports Videos. IEEE Multimedia (2002)
4. Lawrence E. Marks: The Unity of the Senses: Interrelations among the Modalities. New York: Academic Press (1978)
5. Stein, B., Meredith, M.A.: The Merging of the Senses. MIT Press, Cambridge, MA (1993)
6. Gjini, K., Maeno, T., Iramina, K., Ueno, S.: Short-term episodic memory encoding in the human brain: A MEG and EEG study. IEEE Transactions on Magnetics **41**(10) (October 2005) 4149–4151
7. Nakagawa, S., Ueno, S., Imada, T.: Measurements and source estimations of extremely low frequency brain magnetic fields in a short-term memory task by a whole-head neurogradiometer. IEEE Transactions on Magnetics **35**(5) (September 1999) 4130–4132
8. Pick, H.L., Saltzman, E.: Modes of perceiving and processing information. Modes of perceiving and processing information (March 1978) 1–20
9. Bower, T.G.R.: The evolution of sensory systems. Perception: Essays in Honor of James J. Gibson (1974) 141–152
10. Lee, D.: The functions of vision. Modes of perceiving and processing information (March 1978) 159–170
11. Logie, R.: Visuo-spatial working memory. Lawrence Erlbaum Associates (1995)
12. Just, M., Carpenter, P., Hemphill, D.: Constraints on processing capacity: Architectural or implementational. In Steier, D., T.M, M., eds.: Mind matters: A tribute to Allen Newell. Erlbaum (1996)
13. Anderson, J.: The architecture of cognition. Harvard University Press (1983)
14. Dasarathy, B.V.: Sensor fusion potential exploitation – innovative architectures and illustrative applications. Proceedings of the IEEE (1997)
15. Fayman, J.A., Pirjanian, P., henrik I. Christensen, Rivlin, E.: Exploiting process integration and composition in the context of active vision. IEEE Transactions on Systems, Man and Cybernetics — Part C: Applications and Reviews **29**(1) (February 1999) 73–86
16. Bauckhage, C., Wachsmuth, S., Hanheide, M., Wrede, S., Sagerer, G., Heidemann, G., Ritter, H.: The visual active memory perspective on integrated recognition systems. Image and Vision Computing **In Press** (2006)
17. *Authors removed*: A contextual reasoning framework for scene interpretation and tracking in tennis video sequences. Computer Vision and Image Understanding **Submitted for publication** (2006)
18. International Tennis Federation: Rules of Tennis. (2006)

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de