A Self-Referential Perceptual Inference Framework for Video Interpretation

Christopher Town[1] and David Sinclair[2]

[1] University of Cambridge Computer Laboratory, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK cpt23@cam.ac.uk
[2] Waimara Ltd, 115 Ditton Walk, Cambridge UK das@waiamara.com

Abstract. This paper presents an extensible architectural model for general content-based analysis and indexing of video data which can be customised for a given problem domain. Video interpretation is approached as a joint inference problems which can be solved through the use of modern machine learning and probabilistic inference techniques. An important aspect of the work concerns the use of a novel active knowledge representation methodology based on an ontological query language. This representation allows one to pose the problem of video analysis in terms of queries expressed in a visual language incorporating prior hierarchical knowledge of the syntactic and semantic structure of entities, relationships, and events of interest occurring in a video sequence. Perceptual inference then takes place within an ontological domain defined by the structure of the problem and the current goal set.

1 Introduction

The content-based analysis of digital video footage requires methods which will automatically segment video sequences and key frames into image areas corresponding to salient objects (e.g. people, vehicles, background objects, etc.), track these objects in time, and provide a flexible framework for further analysis of their relative motion and interactions.

We argue that these goals are achievable by following the trend in Computer Vision research to depart from strict "bottom-up" or "top-down" hierarchical paradigms and instead place greater emphasis on the mutual interaction between different levels of representation. Moreover, it is argued that an extensible framework for general robust video object segmentation and tracking is best attained by pursuing an inherently flexible "self-referential" approach. Such a system embodies an explicit representation of its own internal state (different sources of knowledge about a video scene) and goals (finding the object-level interpretation which is most likely given this knowledge and the demands of a particular application). The resulting framework can be customised to a particular problem (e.g. tracking human beings from CCTV footage) by integrating the most appropriate low-level (e.g. facial feature extraction) and high-level (e.g. models of human motion) sources of domain-specific knowledge. The system can then be regarded as combining this information at a meta-level to arrive at the most

J.L. Crowley et al. (Eds.): ICVS 2003, LNCS 2626, pp. 54–67, 2003.

likely interpretation (e.g. labelling a block of moving image regions as representing a human body) of the video data given the available information, possibly undergoing several cycles of analysis-integration-conclusion in the process.

In order to make meaningful inferences during this iterative fusion of different sources of knowledge and levels of feature extraction/representation, it is necessary to place such a methodology within the sound theoretical framework afforded by modern probabilistic inference techniques such as the adaptive Bayesian graphical methods known as Dynamic Belief networks. Dynamic Belief networks are particularly suitable because they model the evolution and integration of stochastic state information over time and can be viewed as generalisations of a broad family of probabilistic models.

A key part of the proposed approach concerns the notion that many tasks in computer vision are closely related to, and may be addressed in terms of, operations in language processing. In both cases one ultimately seeks to find symbolic representations which can serve as meaningful interpretations of underlying signal data. Such an analysis needs to incorporate a notion of the syntax and semantics which are seen as governing the domain of interest so that the most likely explanation of the observed data can be found. Whereas speech and language processing techniques are concerned with the analysis of sound patterns, phonemes, words, sentences, and dialogues, video analysis is confronted with pixels, video frames, primitive features, regions, objects, motions, and events. An important difference [32] between the two arises from the fact that visual information is inherently more ambiguous and semantically impoverished. There consequently exists a wide semantic gap between human interpretations of image information and that currently derivable by means of a computer.

We argue that this gap can be narrowed for a particular application domain by means of an ontological language which encompasses a hierarchical representation of task-specific attributes, objects, relations, temporal events, etc., and relates these to the processing modules available for their detection and recognition from the underlying medium. Words in the language therefore carry meaning directly related to the appearance of real world objects. Visual inference tasks can then be carried out by processing sentence structures in an appropriate ontological language. Such sentences are not purely symbolic since they retain a linkage between the symbol and signal levels. They can therefore serve as a computational vehicle for active knowledge representation which permits incremental refinement of alternate hypotheses through the fusion of multiple sources of information and goal-directed feedback to facilitate disambiguation in a context specified by the current set of ontological statements. Particular parts of the ontological language model may be implemented as Dynamic Belief networks, stochastic grammar parsers, or neural networks, but the overall frameworks need not be tied to a particular formalism such as the propagation of conditional probability densities. Later sections will discuss these issues further in light of related work and ongoing research efforts.

2 Related Work

2.1 Visual Recognition as Perceptual Inference

An increasing number of research efforts in medium and high level video analysis can be viewed as following the emerging trend that object recognition and the recognition of temporal events are best approached in terms of generalised language processing which attempts a machine translation [14] from information in the visual domain to symbols and strings composed of predicates, objects, and relations. The general idea is that recognising an object or event requires one to relate ill-defined symbolic representations of concepts to concrete instances of the referenced object or behaviour pattern. This is best approached in a hierarchical manner by associating individual parts at each level of the hierarchy according to rules governing which configurations of the underlying primitives give rise to meaningful patterns at the higher semantic level. Many state-of-the-art recognition systems therefore explicitly or implicitly employ a probabilistic grammar which defines the syntactic rules which can be used to recognise compound objects or events based on the detection of individual components corresponding to detected features in time and space. Recognition then amounts to parsing a stream of basic symbols according to prior probabilities to find the most likely interpretation of the observed data in light of the top-level starting symbols in order to establish correspondence between numerical and symbolic descriptions of information. This idea has a relatively long heritage in syntactic approaches to pattern recognition [39,4] but interest has been revived recently in the video analysis community following the popularity and success of probabilistic methods such as Hidden Markov models (HMM) and related approaches adopted from the speech and language processing community.

While this approach has shown great promise for applications ranging from image retrieval to face detection to visual surveillance, a number of problems remain to be solved. The nature of visual information poses hard challenges which hinder the extent to which mechanisms such as Hidden Markov models and stochastic parsing techniques popular in the speech and language processing community can be applied to information extraction from images and video. Consequently there remains some lack of understanding as to which mechanisms are most suitable for representing and utilising the syntactic and semantic structure of visual information and how such frameworks can best be instantiated. The role of machine learning in computer vision continues to grow and recently there has been a very strong trend towards using Bayesian techniques for learning and inference, especially factorised graphical probabilistic models [23] such as Dynamic Belief networks (DBN). While finding the right structural assumptions and prior probability distributions needed to instantiate such models requires some domain specific insights, Bayesian graphs generally offer greater conceptual transparency than e.g. neural network models since the underlying causal links and prior beliefs are made more explicit. The recent development of various approximation schemes based on iterative parameter variation or stochastic sampling for inference and learning have allowed researchers to construct proba-

bilistic models of sufficient size to integrate multiple sources of information and model complex multi-modal state distributions. Recognition can then be posed as a joint inference problem relying on the integration of multiple (weak) clues to disambiguate and combine evidence in the most suitable context as defined by the top level model structure.

One of the earlier examples of using Dynamic Belief networks (DBN) for visual surveillance appears in [5]. DBNs offer many advantages for tracking tasks such as incorporation of prior knowledge and good modelling ability to represent the dynamic dependencies between parameters involved in a visual interpretation. Their application to multi-modal and data fusion [38] can utilise fusion strategies of e.g. Kalman [10] and particle filtering [20] methods. As illustrated by [11] and [33], concurrent probabilistic integration of multiple complementary and redundant cues can greatly increase the robustness of multi-hypothesis tracking.

In [29] tracking of a person's head and hands is performed using a Bayesian Belief network which deduces the body part positions by fusing colour, motion and coarse intensity measurements with context dependent semantics. Later work by the same authors [30] again shows how multiple sources of evidence (split into necessary and contingent modalities) for object position and identity can be fused in a continuous Bayesian framework together with an observation exclusion mechanism. An approach to visual tracking based on co-inference of multiple modalities is also presented in [41] which describes an sequential Monte Carlo approach to co-infer target object colour, shape, and position. In [7] a joint probability data association filter (JPDAF) is used to compute the HMM's transition probabilities by taking into account correlations between temporally and spatially related measurements.

2.2 Recognition of Actions and Structured Events

Over the last 15 years there has been growing interest within the computer vision and machine learning communities in the problem of analysing human behaviour in video. Such systems typically consist of a low or mid level computer vision system to detect and segment a human being or object of interest, and a higher level interpretation module that classifies motions into atomic behaviours such as hand gestures or vehicle manoeuvres. Higher-level visual analysis of compound events has in recent years been performed on the basis of parsing techniques using a probabilistic grammar formalism. Such methods are capable of recognising fairly complicated behavioural patterns although they remain limited to fairly circumscribed scenarios such as sport events [18,19], small area surveillance [36, 26], and game playing [25]. Earlier work on video recognition such as [40] and [15] already illustrated the power of using a context dependent semantic hierarchy to guide focus of attention and combination of plausible hypothesis, but lacked a robust way of integrating multiple sources of information in a probabilistically sound way.

The role of attentional control for video analysis was also pointed out in [6]. The system described there performs selective processing in response to user

queries for two cellular imaging applications. This gives the system a goal directed attentional control mechanism since the most appropriate visual analysis routines are performed in order to process the user query. Selective visual processing on the basis of Bayes nets and decision theory has also been demonstrated in control tasks for active vision systems [28]. Knowledge representation using Bayesian networks and sequential decision making on the basis of expected cost and utility allow selective vision systems to take advantage of prior knowledge of a domain's cognitive and geometrical structure and the expected performance and cost of visual operators. An interesting two-level approach to parsing actions and events in video is described in [21]. HMMs are used to detect candidate low-level temporal features which are then parsed using a SCFG parsing scheme which adds disambiguation and robustness to the stream of detected atomic symbols. A similar approach is taken by [25] which uses the Earley-Stolcke parsing algorithm for stochastic context-free grammars to determine the most likely semantic derivation for recognition of complex multi-tasked activities from a given video scenario. A method for recognising complex multi-agent action is presented in [19]. Belief networks are again used to probabilistically represent and infer the goals of individual agents and integrate these in time from visual evidence. Bayesian techniques for integrating bottom-up information with top-down feedback have also been applied to challenging tasks involving the recognition of interactions between people in surveillance footage [26]. [24] presents an ontology of actions represented as states and state transitions hierarchically organised from most general to most specific (atomic).

3 Proposed Approach and Methodology

3.1 Overview

We propose a cognitive architectural model for video interpretation. It is based on a self-referential (the system maintains an internal representation of its goals and current hypotheses) probabilistic model for multi-modal integration of evidence (e.g. motion estimators, edge trackers, region classifiers, face detectors, shape models, perceptual grouping operators) and context-dependent inference given a set of representational or derivational goals (e.g. recording movements of people in a surveillance application). The system is capable of maintaining multiple hypotheses at different levels of semantic granularity and can generate an consistent interpretation by evaluating a query expressed in an ontological language. This language gives a probabilistic hierarchical representation incorporating domain specific syntactic and semantic constraints to enable robust analysis of video sequences from a visual language specification tailored to a particular application and for the set of available component modules.

From an Artificial Intelligence point of view this might be regarded as an approach to the *symbol grounding problem* [16] (sentences in the ontological language have an explicit foundation of evidence in the feature domain, so there is a way of bridging the semantic gap between the signal and symbol level) and *frame problem* [12] (there is no need to exhaustively label everything that is going

on, one only needs to consider the subset of the state space required to make a decision given a query which implicitly narrows down the focus of attention).

The nature of such queries will be task specific. They may either be explicitly stated by the user (e.g. in a video retrieval task) or implicitly derived from some notion of the system's goals. For example, a surveillance task may require the system to register the presence of people who enter a scene, track their movements, and trigger an event if they are seen to behave in a manner deemed "suspicious" such as lingering within the camera's field of view or repeatedly returning to the scene over a short time scale. Internally the system could perform these functions by generating and processing queries of the kind "does the observed region movement correspond to a person entering the scene?", "has a person of similar appearance been observed recently?", or "is the person emerging from behind the occluding background object the same person who could no longer be tracked a short while ago?". These queries would be phrased in a language which relates them to the corresponding feature extraction modules (e.g. a Dynamic Belief network for fusing various cues to track people-shaped objects) and internal descriptions (e.g. a log of events relating to people entering or leaving the scene at certain locations and times, along with parameterised models of their visual appearance). Formulating and refining interpretations then amounts to selectively parsing such queries.

3.2 Recognition and Classification

The notion of image and video interpretation relative to the goal of satisfying a structured user query (which may be explicit or implicitly derived from a more general specification of system objectives) follows the trend in recent approaches to robust object recognition on the basis of a "union of weak classifiers". Such an approach hierarchically integrates trained parts-based relationships between lower level feature classifiers to recognise composite objects. Salient perceptual groupings of image features are detected as *non-accidental* image structure identified by means of a particular set of predicates over lower-level image properties (e.g. texture, shape, colour). Making such methods robust, scalable, and generally applicable has proven a major problem.

We argue that in order to come closer to capturing the semantic "essence" of an image or sequence, tasks such as feature grouping and object identification need to be approached in an adaptive goal oriented manner. This takes into account that criteria for what constitutes non-accidental and perceptually significant visual properties necessarily depend on the objectives and prior knowledge of the observer. Such criteria can be ranked in a hierarchy and further divided into those which are *necessary* for the object or action to be recognised and those which are merely *contingent*. Such a ranking makes it possible to quickly eliminate highly improbable or irrelevant configurations and narrow down the search window. The combination of individually weak and ambiguous clues to determine object presence and estimate overall probability of relevance builds on recent approaches to robust object recognition and can be seen as an attempt at extending the success of indicative methods for content representation in the

field of information retrieval. Devising a strategy for recognising objects by applying the most appropriate combination of visual routines such as segmentation and classification modules can also be learned from data [13].

3.3 The Role of Language in Vision

As mentioned above, many problems in vision such as object recognition ([14]), video analysis ([18,27,24]), gesture recognition ([3,21,25]), and multimedia retrieval ([22,2,37]) can be viewed as relating symbolic terms to visual information by utilising syntactic and semantic structure in a manner related to approaches in speech and language processing [34]. A visual language can also serve as an important mechanism for attentional control by constraining the range of plausible feature configurations which need to be considered when performing a visual tasks such as recognition. Processing may then be performed selectively in response to queries formulated in terms of the structure of the domain, i.e. relating high-level symbolic representations to extracted features in the signal (image and temporal feature) domain. By basing such a language on an ontology one can capture both concrete and abstract relationships between salient visual properties. Ontologies encode the relational structure of concepts which one can use to describe and reason about aspects of the world.

Since the language is used to express queries and candidate hypotheses rather than exhaustively label image content, such relationships can be represented explicitly without prior commitments to a particular interpretation or having to incur the combinatorial explosion of a full annotation of all the relations that may hold in a given image or video. Instead, only those image aspects which are of value given a particular query are evaluated and evaluation may stop as soon as the appropriate top level symbol sequence has been generated.

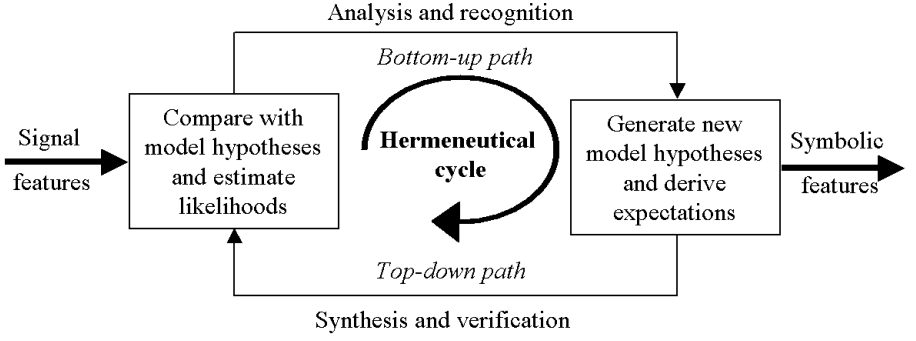

Fig. 1. The Hermeneutical cycle for iterative interpretation in a generative (hypothesise and test) framework.

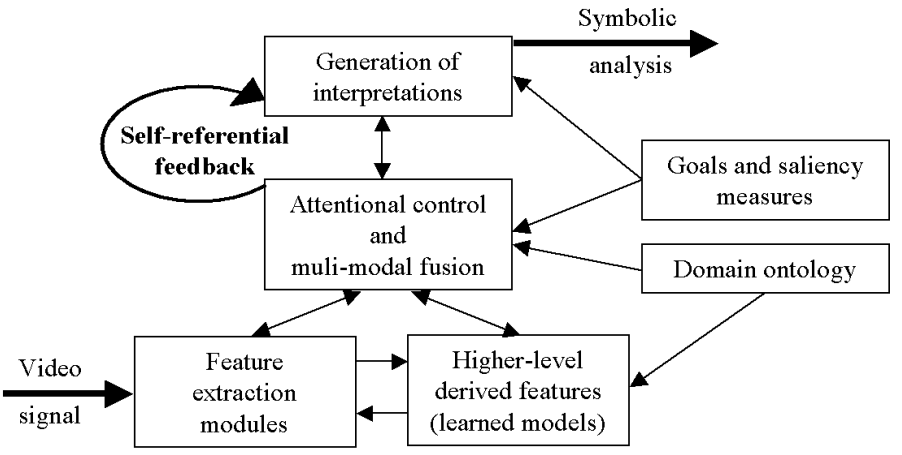

Fig. 2. Sketch of the proposed approach to goal-directed fusion of content extraction modules and inference guided by an attentional control mechanism. The fusion process and selective visual processing are carried out in response to a task and domain definition expressed in terms of an ontological language. Interpretations are generated and refined by deriving queries from the goals and current internal state.

3.4 Self-Referential Perceptual Inference Framework

In spite of the benefits of DBNs and related formalisms outlined above, probabilistic graphical models also have limitations in terms of their ability to represent structured data at a more symbolic level and the requirement for normalisations to enable probabilistic interpretations of information. Devising a probabilistic model is in itself not enough since one requires a framework which determines which inferences are actually made and how probabilistic outputs are to be interpreted.

Interpreting visual information in a dynamic context is best approached as an iterative process where low-level detections are compared (induction) with high-level models to derive new hypotheses (deduction). These can in turn guide the search for evidence to confirm or reject the hypotheses on the basis of expectations defined over the lower level features. Such a process is well suited to a generative method where new candidate interpretations are tested and refined over time. Figure 1 illustrates this approach.

However, there is a need to improve on this methodology when the complexity of the desired analysis increases, particularly as one considers hierarchical and interacting object and behavioural descriptions best defined in terms of a syntax at the symbolic level. The sheer number of possible candidate interpretations and potential derivations soon requires a means of greatly limiting the system's focus of attention. A useful analogy is selective processing in response to queries [6]. Visual search guided by a query posed in a language embodying an ontological

representation of a domain allows adaptive processing strategies to be utilised and gives an effective attentional control mechanism.

We argue that an ontological content representation and query language can be used as an effective vehicle for hierarchical representation and goal-directed inference in video analysis tasks. As sketched in figure 2, such a language serves as a means of guiding the fusion of multiple sources of visual evidence and refining symbolic interpretations of dynamic scenes in the context of a particular task. By maintaining representations of both the current internal state and derivational goals expressed in terms of the same language framework, such a system can be seen as performing self-referential feedback based control of the way in which information is processed over time. Visual recognition then amounts to selecting a parsing strategy which determines how elements of the current string set are to be processed further given a stream of lower level tokens generated by feature detectors. Parts of the language may be realised in terms of probabilistic fusion mechanisms such as DBNs, but the overall structure of the interpretative module is not limited to a particular probabilistic framework and allow context-sensitive parsing strategies to be employed where appropriate.

4 Applications

4.1 Image and Video Indexing

In [37] we proposed an ontological query language called OQUEL as a novel query specification interface and retrieval tool for content based image retrieval and presented results using the ICON system. The language features an extensible language framework based on a formally specified grammar and vocabulary which are derived from a general ontology of image content in terms of categories, objects, attributes, and relations. Words in the language represent predicates on image features and target content at different semantic levels. Sentences are prescriptions of desired characteristics which are to hold for relevant retrieved images. Images are retrieved by deriving an abstract syntax tree from a textual or forms-based user query and probabilistically evaluating it by analysing the composition and perceptual properties of salient image regions in light of the query. The matching process utilises automatically extracted image segmentation and classification information and can incorporate any other feature extraction mechanisms or contextual knowledge available at processing time to satisfy a given user request. Perceptual inference takes the form of identifying those images as relevant for which one can find sufficient support for the candidate hypotheses derived from the query relative to other images in the collection. Examples of queries are *"some sky which is close to buildings in upper corner, size at least 20%"* and *"(some green or vividly coloured vegetation in the centre) which is of similar size as (clouds or blue sky at the top)"*.

The OQUEL language is currently being extended to the video domain for indexing purposes. This work employs the region based motion segmentation method described in [31] which uses a Bayesian framework to determined the most likely labelling of regions according to motion layers and their depth ordering. The inference framework described above is then utilised to integrate

information form the neural network region classifiers to modify the prior probabilities for foreground/background layer assignments of image regions. A face detector and simple human shape model have recently been used to identify and track people. An ontological language is under development which extends the static scene content descriptions with motion verbs ("moves", "gestures"), spatial and temporal prepositions ("on top of", "beside", "before"), and adverbs ("quickly", "soon") for indexing and retrieval of video fragments.

4.2 Multi-modal Fusion for Sentient Computing

Interesting avenues for refinement, testing and deployment of the proposed cognitive inference framework arise from the "sentient computing" ([17,1]) project developed at AT&T Laboratories Cambridge and the Cambridge University Laboratory for Communications Engineering (LCE). This system uses mobile ultrasonic sensor devices known as "bats" and a receiver infrastructure to gather high-resolution location information for tagged objects such as people and machines in order to maintain a sophisticated software model of an office environment. Applications can register with the system to receive notifications of relevant events to provide them with an awareness of the spatial context in which users interact with the system. As indicated in figures 3 and 4, the system's internal dynamic representation is based on an ontology in terms of locations and spatial regions, objects (people, computers, phones, devices, cameras, furniture etc.), and event states (motions, spatial overlap, proximity, button events etc.).

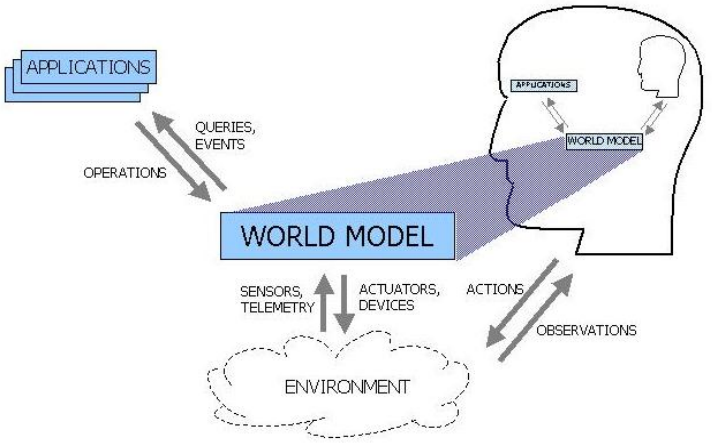

Fig. 3. Diagrammatic overview of the world model maintained by the sentient computing system.

Combining vision with other sensory modalities is a very promising research avenue in ubiquitous perceiving systems [8,35,9]. Computer vision methods can

Fig. 4. The world as perceived by (a) users and (b) the sentient computing system.

provide multi-modal human-computer interfaces with transparent detection, recognition, and tracking capabilities, but on their own suffer from a lack of robustness and autonomy in real world interaction scenarios. The sentient computing system provides a variable granularity spatial model of the environment and a reliable device tracking facility which can be used to automatically (re)initialise and re-focus vision modules whenever an event or scene context of interest is observed by a camera. Location events from the sentient computing architecture are being used to provide ground truth information about objects and object movements occurring within the known field of view of a calibrated camera to yield training and test data for individual video analysis components. A number of applications of the sentient computing technology can in turn benefit from our proposed video interpretation framework through the fusion of the ultrasonic and visual modalities. One such application currently under development concerns user authentication for security critical applications, e.g. those which allow users to automatically unlock office doors or automatically login to a computer in their vicinity. In this case the system uses the bat sensor information to detect that a person is present and select one or more camera views to verify their identity as indicated by the identity tag of their bat. The location of the bat is used to constrain the search window for a head and face detector which forwards an image of the detected face to a face recogniser. Rather than solving the extremely difficult problem of general face recognition, visual authentication is approached as a verification problem and greatly constrained by fusing other kinds of information about assumed identity, face location, lighting conditions, and local office geometry.

Having visual information as an additional sensory modality is also useful when the system has trouble detecting a person (e.g. they are not wearing a bat or it is temporarily concealed) or when an application requires additional information about a person's posture, direction of gaze, gestures, interactions with devices and other people, or facial expression to enhance visually mediated human computer interaction and provide a richer model of the context in which such interactions take place. At a more mundane level, vision technology makes the installation, maintenance and operation of a sentient computing sys-

tem easier by providing additional means of calibrating sensory infrastructure and adapting a model of the static environment (such as furniture and partition walls).

By ensuring that the symbolic inferences drawn by the system remain grounded in the signal domain, the system can support a range of possible queries as inferences and adapt its hypotheses in light of new evidence. To ensure sufficient performance to enable real-time processing, the fusion of individual perceptual modalities is set up as a hierarchy where inexpensive detectors (e.g. finding the rough outline of a person) narrow down the search space to which more specific modules (e.g. a face spotter or gesture recogniser) are applied. The system thereby remains robust to error rates by integrating information vertically (applying detectors with high false acceptance rates to guide those with potentially high false rejection rates) and horizontally (fusing different kinds of information at the same level to offset different error characteristics for disambiguation). In our cognitive framework, vision is therefore used to enhance the perceptual inference capabilities of the sentient computing infrastructure by adding further sources of information to update, query, and extend the system's internal ontology and external event model. By maintaining a notion of its own internal state and goals the system can restrict its focus of attention to perform only those inferences which are required for the current task (e.g. verifying the identity of a person who just entered the visual field). Real-time requirements and other resource limitations can be used as additional constraints for the fusion process.

5 Conclusion

This paper presents an extensible video analysis framework which can be customised for a given task domain by employing appropriate data sources and application-specific constraints. Recent advances in graph-based probabilistic inference techniques allow the system to propagate a stochastic model over time and combine different types of syntactic and semantic information. The process of generating high-level interpretations subject to system goals is performed by parsing sentence forms in an ontological language for visual content at different levels of analysis.

Acknowledgements. The authors would like to acknowledge directional guidance and support from AT&T Laboratories and the Cambridge University Laboratory for Communications Engineering. The principal author received financial support from the Royal Commission for the Exhibition of 1851.

References

1. M. Addlesee, R. Curwen, S. Hodges, J. Newman, P. Steggles, A. Ward, and A. Hopper. Implementing a sentient computing system. *IEEE Computer*, 34(8):50–56, 2001.
2. K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proc. International Conference on Computer Vision*, 2001.

3. A. Bobick and Y. Ivanov. Action recognition using probabilistic parsing. In *Proc. Conference on Computer Vision and Pattern Recognition*, 1998.
4. H. Bunke and D. Pasche. *Structural Pattern Analysis*, chapter Parsing multivalued strings and its application to image and waveform recognition. World Scientific Publishing, 1990.
5. H. Buxton and S. Gong. Advanced visual surveillance using bayesian networks. In *Proc. International Conference on Computer Vision*, 1995.
6. H. Buxton and N. Walker. Query based visual analysis: Spatio-temporal reasoning in computer vision. *Vision Computing*, 6(4):247–254, 1988.
7. Y. Chen, Y. Rui, and T. Huang. JPDAF based HMM for real-time contour tracking. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2001.
8. J. Crowley, J. Coutaz, and F. Berard. Things that see: Machine perception for human computer interaction. *Communications of the ACM*, 43(3):54–64, 2000.
9. J. Crowley, J. Coutaz, G. Rey, and P. Reignier. Perceptual components for context aware computing. In *Proc. Ubicomp 2002*, 2002.
10. J. Crowley and Y. Demazeau. Principles and techniques for sensor data fusion. *Signal Processing*, 32(1–2):5–27, 1993.
11. T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. In *Proc. Conference on Computer Vision and Pattern Recognition*, 1998.
12. D. C. Dennett. *Minds, machines, and evolution*, chapter Cognitive Wheels: The Frame Problem of AI, pages 129–151. Cambridge University Press, 1984.
13. B. Draper, U. Ahlrichs, and D. Paulus. Adapting object recognition across domains: A demonstration. *Lecture Notes in Computer Science*, 2095:256–270, 2001.
14. P. Duygulu, K. Barnard, J.F.H. De Freitas, and D.A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. European Conference on Computer Vision*, 2002.
15. J. Glicksman. A cooperative scheme for image understanding using multiple sources of information. Technical Report TR-82-13, University of British Columbia, Department of Computer Science, 1982.
16. S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.
17. A. Harter, A. Hopper, P. Steggles, A. Ward, and P. Webster. The anatomy of a context-aware application. In *Mobile Computing and Networking*, pages 59–68, 1999.
18. G. Herzog and K. Rohr. Integrating vision and language: Towards automatic description of human movements. In I. Wachsmuth, C.-R. Rollinger, and W. Brauer, editors, *KI-95: Advances in Artificial Intelligence. 19th Annual German Conference on Artificial Intelligence*, pages 257–268. Springer, 1995.
19. S. Intille and A. Bobick. Representation and visual recognition of complex, multi-agent actions using belief networks. In *IEEE Workshop on the Interpretation of Visual Motion*, 1998.
20. M. Isard and A. Blake. ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. *Lecture Notes in Computer Science*, 1406, 1998.
21. Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 22(8), 2000.
22. A. Jaimes and S. Chang. A conceptual framework for indexing visual information at multiple levels. In *IS&T SPIE Internet Imaging*, 2000.
23. F.V. Jensen. *An Introduction to Bayesian Networks.* Springer Verlag, 1996.

24. A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. Int. Journal of Computer Vision (to appear), 2002.
25. D. Moore and I. Essa. Recognizing multitasked activities using stochastic context-free grammar. In *Proc. Workshop on Models vs Exemplars in Computer Vision*, 2001.
26. N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 22(8):831–843, 2000.
27. C. Pinhanez and A. Bobick. Approximate world models: Incorporating qualitative and linguistic information into vision systems. In *AAAI'96*, 1996.
28. R. Rimey. *Control of Selective Perception using Bayes Nets and Decision Theory.* PhD thesis, University of Rochester Computer Science Department, 1993.
29. J. Sherrah and S. Gong. Tracking discontinuous motion using bayesian inference. In *Proc. European Conference on Computer Vision*, pages 150–166, 2000.
30. J. Sherrah and S. Gong. Continuous global evidence-based bayesian modality fusion for simultaneous tracking of multiple objects. In *Proc. International Conference on Computer Vision*, 2001.
31. P. Smith. *Edge-based Motion Segmentation.* PhD thesis, Cambridge University Engineering Department, 2001.
32. K. Sparck Jones. Information retrieval and artificial intelligence. *Artificial Intelligence*, 114:257–281, 1999.
33. M. Spengler and B. Schiele. Towards robust multi-cue integration for visual tracking. *Lecture Notes in Computer Science*, 2095:93–106, 2001.
34. R. Srihari. Computational models for integrating linguistic and visual information: A survey. *Artificial Intelligence Review, special issue on Integrating Language and Vision*, 8:349–369, 1995.
35. S. Stillman and I. Essa. Towards reliable multimodal sensing in aware environments. In *Proc. Perceptual User Interfaces Workshop, ACM UIST 2001*, 2001.
36. M. Thonnat and N. Rota. Image understanding for visual surveillance applications. In *Proc. of 3rd Int. Workshop on Cooperative Distributed Vision*, 1999.
37. C.P. Town and D.A. Sinclair. Ontological query language for content based image retrieval. In *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, pages 75–81, 2001.
38. K. Toyama and E. Horvitz. Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *Proc. Asian Conference on Computer Vision*, 2000.
39. W. Tsai and K. Fu. Attributed grammars - a tool for combining syntactic and statistical approaches to pattern recognition. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-10(12), 1980.
40. J. Tsotsos, J. Mylopoulos, H. Covvey, and S. Zucker. A framework for visual motion understanding. *IEEE Trans. on Pattern Analysis and Machine Intell.*, Special Issue on Computer Analysis of Time-Varying Imagery:563–573, 1980.
41. Y. Wu and T. Huang. A co-inference approach to robust visual tracking. In *Proc. International Conference on Computer Vision*, 2001.