A Lightweight Camera Sensor Network Operating on Symbolic Information

Thiago Teixeira *, Dimitrios Lymberopoulos *,
Eugenio Culurciello *, Yiannis Aloimonos†, and Andreas Savvides *
*Dept. of Electrical Engineering
Yale University, New Haven, CT 06520 USA
Email: {thiago.teixeira, dimitrios.lymberopoulos, eugenio.culurciello, andreas.savvides}@yale.edu
†Computer Vision Laboratory, Dept. of Computer Science
University of Maryland, College Park, MD 20742 USA
Email: {yiannis}@cs.umd.com

***Abstract*— This paper provides an overview of the research aspects of our DSC06 demonstration. We present a new camera sensor network for behavior recognition. Two new technologies are explored, biologically inspired address-event image sensors and sensory grammars. This paper explains how these two technologies are used together and reports of the current status of our prototyping effort. The application of the resulting system in assisted living is also described.**

I. Introduction

The infusion of wireless sensor networks to everyday life situations is reviving the interest in the creation of new lightweight camera sensor networks. Camera sensors, perhaps the most information-rich sensing modality, are becoming smaller, low-power and more affordable. With such changes in technology, their deployment in large numbers for higher fidelity data makes a lot of sense, but also poses a new set of challenges. Camera networks are expected to operate over low-bandwidth links, self-calibrate and consume little power. Also despite the need for accurate information, the lack of privacy preservation in cameras makes people unconfortable using them and often raises legal issues related to privacy.

Our work tries to leverage the higher information quality of cameras and imagers while trying to address some of the aforementioned challenges. We do so by working towards the creation of a lightweight network of camera sensors that operates on symbolic information rather than images. To minimize power consumption and bandwidth, and to mitigate privacy concerns we pursue the development of a new generation of biologically-inspired image sensors. These specialize in picking only the useful information from a scene to reduce processing and bandwidth, and supply their outputs in an address-event stream that is inherently more privacy-preserving. To avoid the communication of raw data over wireless links, we are also working towards the development of a behavior interpretation framework based on a hierarchy of probabilistic grammars that can convert low-level sensor measurements to higher-level behavior interpretation. The combination of these two technologies aim to create a network that can efficiently convert signals to semantics at the node level and can perform robust behavior interpretation at the network level, by operating on symbolic information.

In this demo paper, we summarize different aspects of our work, and our efforts towards the deployment of the resulting system in an assisted living application. Our presentation begins with an overview of our two core technologies: custom address-event image sensors and sensory grammars. We then describe our prototype platforms and software services, and comment on power consumption observations. The paper concludes with a brief description of the application of our network in assisted living.

II. Towards Lightweight Camera Networks that Operate on Symbolic Information

A large component of our work tries to address a gap in technology for sensing motion, particularly that of humans. Today, human motion can be detected with Passive Infrared Sensors (PIR) but cannot be accurately observed without the use of cameras. Cameras, however, require extensive resources in terms of computation, memory and communication. Our work tries to define a new, motion-discriminating, sensing modality defined by hardware and software. Instead of giving image outputs, this new modality outputs symbols that summarize motion activity. In hardware, the sensing will be performed with an array of pixels in a custom imager architecture that filters the visual scene to provide numerical outputs to the node processor. These outputs will then be processed directly by a sensory grammar hierarchy into semantic form before it is transmitted inside the network for more complex behavior identification. An overview of address-event image sensors and sensory grammars is given below.

A. Address Event Image Sensors

AER is, strictly speaking, a communication protocol for biomimetic chips (Figure 1). It was developed to permit the interaction between independently-designed neural-network ICs. In AER systems, an *address* is assigned to each *event* detected by a sensor. The name *address* comes from neural-networks and traditionally corresponds to the address of the neuron that detected the event in question. In effect, this address can be thought of as a *description* of the event, and can be directly

converted to a grammar symbol. Thus, AER sensors can be easily connected into the sensing grammars without the need for additional processing. AER imagers do not communicate pixel values explicitly, like traditional cameras. Instead, they produce a stream of addresses consisting of the addresses of the pixels that met a certain criteria. Information about the *intensity* of each event is typically contained in the frequency of recurring addresses. The exact criteria utilized for triggering events varies according to the needs of the application. In the ALOHA imager [4], for example, each pixel triggers an event whenever it has acquired a pre-defined amount of light. On some of the imagers we simulate (as discussed below), we utilize the presence of temporal differences or even spatial edges as a criteria for triggering an event at a pixel.

For this reason, AER imagers have the potential to act as "blind" cameras, which cannot take pictures: instead these imagers *measure* a scene, looking for the most relevant data for the processing job at hand. This is an invaluable tool for applications such as security and assisted living, where privacy preservation is a major concern.

To reduce the design-fabricate-test-redesign cycle and to experiment with different AER imaging technologies in the context of WSNs, an AER imaging emulator for the PC was developed [17]. The software provides a unified interface where multiple parameters can be tweaked. Figure 2 shows three different address-event functions of the emulator: motion, edge and centroid detection. These are accomplished through traditional computer-vision techniques followed by a conversion to address-space, as shown in Figure 3. The output events can then be directly routed to the WSN through TCP sockets for in-node processing. Note that while the original software run on the PC, nowadays emulation happens at the node-level. Later, once the best parameters are found for a particular application, the emulator can be used to guide the design of custom AER image sensors for WSN deployment.

A custom AER image sensor for use in our camera-network is currently in design phase. Once completed, this sensor will allow for much faster data processing since there will be no need for software-based feature detection. This means that it will replace the AER emulator that currently runs on the node and directly communicate with the bottom layers of our existing grammar hierarchies. In the meantime, we utilize the in-node emulation for our deployment.

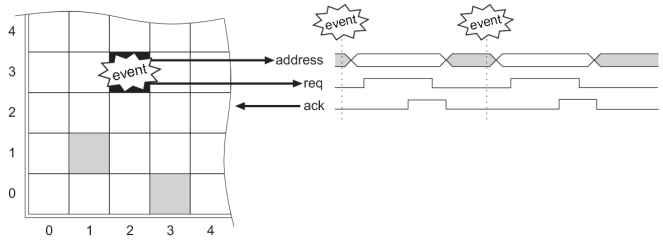

Fig. 1. Visual description of the AER communication protocol. An address is transmitted for each event detected by a sensor.

B. Sensory Grammars

Our approach towards behavior interpretation, is based on the fact that behaviors, particularly those of humans, are patterns in space and time. Depending on the granularity of space and time at which these patterns take place they can be classified as *microscopic* and *macroscopic* patterns. Microscopic patterns take place over the same location and over very small time periods. They are called *human gestures* and they constitute a large field in computer vision research[11], [18], [6], [2], [19]. Macroscopic patterns take place over much larger spaces (room, house, city, etc.) and much larger time intervals (minutes, hours, days, etc). In our project, the primary focus is to take advantage of the distributed nature and scalability of sensor networks to enable the detection of these macroscopic behaviors. Sensor networks today can enable the monitoring of humans over large spaces and time intervals. However what sensor networks cannot do today is to use this monitoring information to *reason* about what humans do. Our intention is to use simple location or area information to reason about macroscopic human behaviors. This type of information can be acquired from a calibrated camera sensor network [3], or another positioning technology that combines basic human motion information with building/city maps to extract the area at which human activity is taking place. More precise sensing of a human's location/area can also be acquired by recording the interaction of humans with several objects using RFID technology as demonstrated in [7], [13], [16], [12], [8].

These sequences of basic sensing features of human activity are fed into a powerful inference engine that translates them into high level human behaviors. As the basis of this inference engine we use probabilistic context free grammars

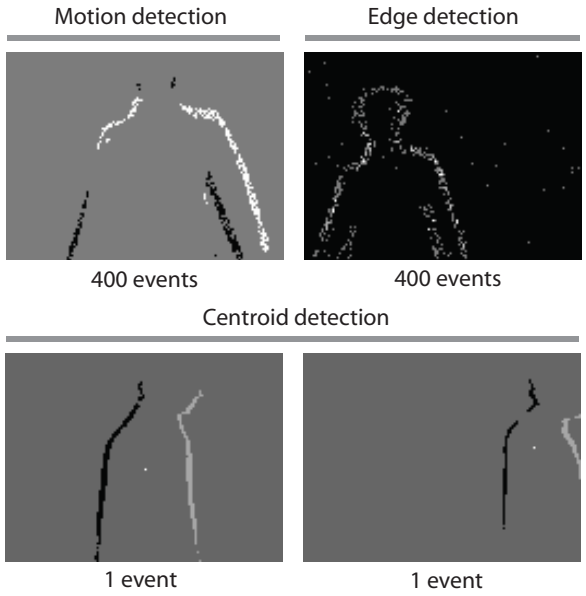

Fig. 2. Output of AER Emulator given a 128×128 input. Only 400 events where utilized for displaying the motion and edge-detected images. On the centroid images, the centroid is the bright pixel within the moving silhouette (1 event).

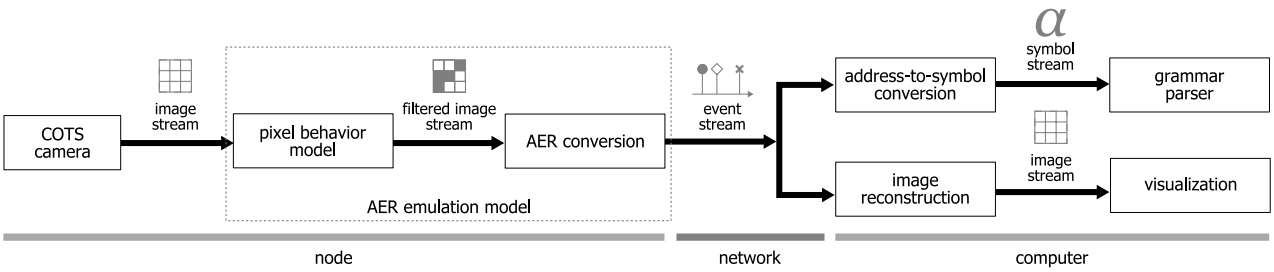

Fig. 3. Block diagram of AER Emulator. Video is acquired by COTS camera, then processed using standard computer-vision techniques. The output is then converted to a stream of events and used for posterior processing in address-space. Alternatively, for visualisation purposes, the event stream can be converted back to video.

(PCFGs)[10], [5], [20], [9]. PCFGs are very similar to the human and computer grammars we are accustomed to. The only difference is that each production rule of the grammar is associated with a probability allowing one to compute a likelihood probability for each output string. PCFGs are very similar to (and often interchangeable with) the Hidden Markov Models (HMMs) used in speech processing and handwriting recognition[14]. What makes them more appealing for use in sensor networks is their expressiveness, generative power and modularity. Using only a few lines of grammar description a large number of different symbol sequences can be described.

The power of this inference framework comes from its hierarchical organization. Once defined, each grammar specification can be viewed as a black box with well-specified inputs or outputs. This makes it easy to compose grammar hierarchies for interpreting raw data by wiring grammars together. Each level in the hierarchy interprets and summarizes its input, providing different interpretations of the same raw data, by focusing on different levels of granularity. This results in a powerful and scalable inference engine in which different applications may choose to make different interpretations of the same underlying sensor data. An example of a grammar hierarchy organization is shown in Figure 4.

Using a hierarchical inference engine we also aim to avoid exhaustive training of the entire sensor network for all possible behaviors and all possible instances of behaviors. By confining the required training at the lowest level of the hierarchy we aim to have behavior independent training. As long as the sensors (or sensor preprocessors) can be trained to output a set of phonemes (location, area, direction, etc.), one could reason about more macroscopic behaviors without requiring further training. Communication and node coordination should be organized in a way such that the recognition of an action or behavior is elastic. The network will bind together the sensor nodes in a way that multiple nodes can achieve the similar quality of recognition to a single sensor with global view, even if the phenomenon is observed by a number of nodes over space and time.

The sensory grammars considered in this project will be used to combine information from multiple sensors (i.e imagers, light, temprature, PIR, acoustic and other custom sensors) in a new way. Existing mathematical techniques try to fuse multimodal sensor measurements by fusing different

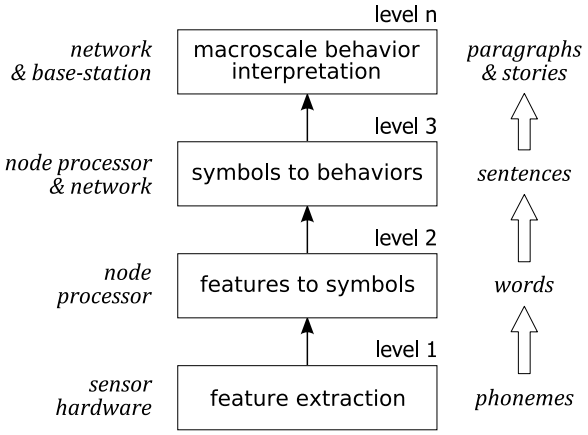

Fig. 4.

probability distributions, a process that is not yet 100% understood. In contrast to these approaches, our approach will take a more pragmatic standpoint depending on the problem and the type of available data. We will exploit this to restrict our search space by looking for combinations of patterns that narrow down a specific activity, and possibly do vision search in this space to refine the result.

An example of a turning sensor based on sensory grammars can be found in [10]. An advanced sensor that can recognize a more complex cooking activity by reasoning on locations and building layout information can be found in [9].

III. Prototype Platforms

A. Platform Direction

The combination of address-event imagers and sensory grammars targets the development of a new low power sensor node architecture that will be able to operate an imager with an 8-bit microcontroller. Despite its low power and limited computation resources, this architecture is expected to carry out the same sensing and classification tasks that are currently only available on higher end nodes, such as the iMote2 based camera sensor node we describe in the next subsection. This is possible because the imager pre-filters the visual scene a provides the data in a format that reduces the need for more elaborate processing. A lifetime comparison as a function of the event arrival rate between the custom imager platforms and

our existing functional COTS prototype is shown in Figure 5. Additional gains are expected from the information reduction taking place in the grammars.

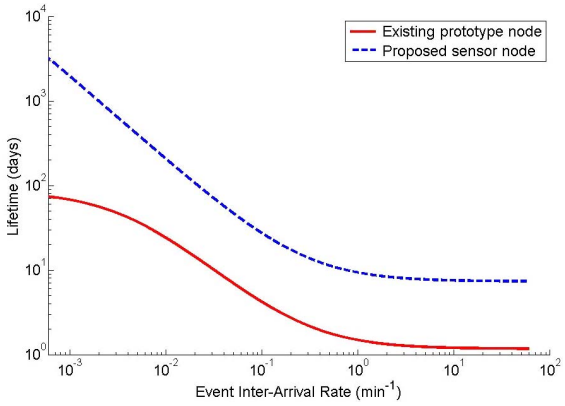

Fig. 5. Lifetime projection for integrated platform

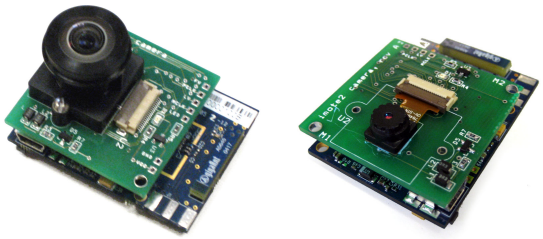

Fig. 6. The two sensor-node configurations: standard (right) and wide-angle (left).

B. Our current functional prototype node

Our camera network utilizes the iMote2 sensor node and a custom camera-board. The iMote2 is made by Intel and bundles a low-power XScale processor (the PXA271) and a 2 GHz 802.15.4 radio from ChipCon (CC2420). The frequency and voltage of the PXA is dynamically scalable (13 MHz to 416 MHz), and there are five major power modes. At deep-sleep, the iMote2 consumes 1.8 mW of power. The sensor-node provides 256 KB of integrated SRAM, 32 MB of external SDRAM, and 32 MB of Strataflash memory.

The camera-board packs an OmniVision OV7649 camera, which can capture color images at 30 fps VGA (640×480) and 60 fps QVGA (320×240). Currently, there are 2 different lens configurations: standard and 162° wide-angle (Figure 6). The power consumption of the active camera-board is of 44 mW. In fully-active mode, at 104 MHz and 8 fps, the entire system consumes 322 mW (of which the iMote2 is responsible for 279 mW).

To locate people in a room, a node placed on the ceiling performs the following operations: first, it acquires an image, then downsamples it to 80×60, compares the result to the previous frame for motion detection, and finally runs the AER Emulator's centroid computation algorithm (Section II). The node then time-stamps each centroid with the value of its real-time clock. Centroids are packed together in packets before being transmitted through the radio, in order to minimize the energy-per-bit. This entire process occurs at a rate of around 8 fps.

Centroids are converted into grammar symbols (level 2 in Figure 4) by producing a different symbol each time a centroid is within one of the predefined areas of interest. Presently, this is done as a post-processing step on the PC, which also runs the grammar parser. In the future, these tasks will take place inside the sensor-node, and only the high-level output of the grammars (low bit-rate) will need to be communicated.

We have also implemented an in-node image-processing library that provides the traditional tools for image manipulation, such as cropping, resampling, colorspace conversion, thresholding, temporal differencing, Sobel edge detection and convolution.

IV. Application to Assisted Living

Our main application focus for this network is assisted living and helping elders living alone. Using a lightweight, privacy preserving network similar to the one described here we are currently developing a system for observing activity inside a house. Our behavior interpretation framework is programmed to recognize unsafe and out-of-the-ordinary behaviors of the inhabitants. Two types of patterns are observed. The first are well defined activities and rules that raise exceptions in our system. The second type is based on longer term statistical properties of behavior. These are meant to recognize shifts in behavior patterns over a time period.

In our current assisted-living deployment, our network recognizes a set of behaviors and rules by reasoning using areas and locations. Seven iMote2 camera nodes featuring the 162° lenses cover the entire floorplan of a two-bedroom apartment. The nodes are deployed at the center of the ceiling in each room, facing down. The nodes measure the location of people in the room, and forward time-stamped locations to a base-station PC. There, the centroids are either stored in an SQL database for posterior analysis or sent to the grammar parsers for interpretation.

Figure 7 shows the trace of a person cooking in a kitchen. It was acquired with the deployment described in [9]. As reported in that paper, the grammar rules enabled the correct identification of cooking activity. It was also able to distinguish "cooking" from "cleaning" actions. An expanded version of this grammar, utilising the full-house deployment and an extended set of activities, is currently in development.

V. Conclusion

Although our work is still in initial stages, the results are very encouraging. A prototype network based on the iMote2 camera nodes is already under deployment in our assisted living application and a detailed set of sensory grammar libraries is under development. The use of sensory grammars for parsing behaviors is not limited to assisted-living. Applications areas from security to gaming are potential targets

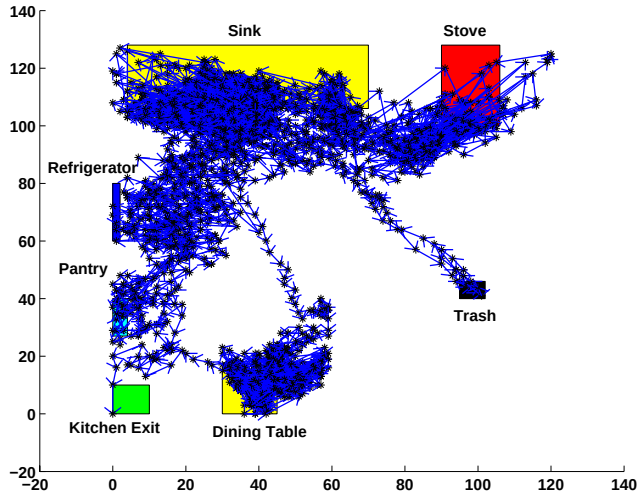

Fig. 7. Experimental trace of a person cooking dinner. This was correctly identified by the grammar hierarchy as "cooking" activity.

for our framework. As a simple demonstration of this, we have developed an augmented-reality game where the user controls a remote-controlled car on the street-map of a city, projected on the floor [15]. The AER Emulator is utilized for tracking the car while its behavior is observed with grammars. The user loses points for each infraction committed, such as driving in the wrong direction. Interestingly, this application can easily be extrapolated from a simple game to a real-life scenario, simply by employing a GPS instead of the address-event sensor. The grammar hierarchy would not require any change to accommodate this.

Additionally, we are working on utilizing AER to summarize image information for a more cost-effective use of radio transmission. In some scenarios, it is interesting to roughly assess the degree of importance of an occurrence before switching into a high-power mode. The type of summaries provided by motion- and edge-detection AER streams (Figure 2) can be of invaluable importance in such scenarios. Instead of constantly transmitting the entire 128×128 15 fps video (over 1.9 Mbps), one could limit the event-rate to a maximum of 6000 events per second (84kbps) – for orders of magnitude of savings! This would allow the reconstruction of 15 400-event images per second, such as the ones in Figure 2. Given this type of summarized event stream, a human agent could then choose whether to commence the high-power video transmission or to wait for a more important occurrence.

For more details and updates on this work please refer to the BehaviorScope Project website [1].

ACKNOWLEDGMENT

The authors would like to acknowledge the help of Andrew Barton-Sweeney and Deokwoo Jung for the camera prototypes and power analysis work. We are also thankful to Lama Nachman of Intel for her help with the iMote2 platform.

REFERENCES

[1] Behaviorscope project website. `http://www.eng.yale.edu/enalab/behaviorscope.htm`.

[2] J. Aggarwal and S. Park. Human motion: Modeling and recognition of actions and interactions. *Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization and Transmission*, pages 640–647, 2004.

[3] A. Barton-Sweeney, D. Lymberopoulos, and A. Savvides. Sensor localization and camera calibration in distributed camera sensor networks. In *Proceedings of IEEE Basenets, October 2006*, October 2006.

[4] E. Culurciello and A. G. Andreou. ALOHA CMOS imager. In *Proceedings of the 2004 IEEE International Symposium on Circuits and SystemsISCAS '04*, May 2004.

[5] S. Geman and M. Johnson. Probabilistic grammars and their applications. In *International Encyclopedia of the Social & Behavioral Sciences. N.J. Smelser and P.B. Baltes, eds., Pergamon, Oxford, 12075-12082*, 2002.

[6] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics Part C*, 34(3):334–352, August 2004.

[7] S. Intille, K. Larson, and E. M. Tapia. Designing and evaluating technology for independent aging in home. In *International Conference oon Aging, Disability and Independence*, 2003.

[8] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition using relational markov models. In *Nineteenth International Joint Conference on Artificial Intelligence*, 2005.

[9] D. Lymberopoulos, A. Barton-Sweeney, T. Teixeira, and A. Savvides. An easy-to-program system for parsing human activities. In *ENALAB Technical Report 060901*, September 2006.

[10] D. Lymberopoulos, A. Ogale, A. Savvides, and Y. Aloimonos. A sensory grammar for inferring behaviors in sensor networks. In *Proceedings of Information Processing in Sensor Networks (IPSN)*, April 2006.

[11] A. Ogale, A. Karapurkar, and Y. Aloimonos. View-invariant modeling and recognition of human actions using grammars. *Workshop on Dynamical Vision at ICCV'05*, October 2005.

[12] D. Patterson and M. P. D. Fox, H. Kautz. Fine-grained activity recognition by aggregating abstract object usage. In *IEEE International Symposium on Wearable Computers*, October 2005.

[13] M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. Kautz, and D. Hahnel. Inferring activities from interactions with objects. *IEEE Pervasive Computing*, 03(4):50–57, 2004.

[14] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proc. IEEE, Vol.77, No. 2, pp. 257-286*, February 1989.

[15] J. Schwarz. A grammar-based system for game playing with a sensor network. `http://www.eng.yale.edu/enalab/publications/Jonathan_grammars.pdf`, May 2006.

[16] E. M. Tapia, S. S. Intille, and K. Larson. Activity recognition in the home setting using simple and ubiquitous sensors. In *PERVASIVE 2004*, 2004.

[17] T. Teixeira, E. Culurciello, J. Park, D. Lymberopoulos, A. Barton-Sweeney, and A. Savvides. Address-event imagers for sensor networks: Evaluation and programming. In *Proceedings of Information Processing in Sensor Networks (IPSN)*, April 2006.

[18] M. Valera and S. Velastin. Intelligent distributed surveillance systems: a review. *IEE Proceedings on Vision, Image and Signal Processing*, 152(2):192–204, April 2005.

[19] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern recognition*, 36:585–601, 2003.

[20] C. S. Wetherell. Probabilistic languages: A review and some open questions. *ACM Comput. Surv.*, 12(4):361–379, 1980.