Use of Context in Vision Processing: An Introduction to the UCVP 2009 Workshop

Hamid Aghajan
Stanford University
Stanford, USA
hamid@wsnl.stanford.edu

Ralph Braspenning
Philips Research
Eindhoven, The Netherlands
ralph.braspenning@philips.com

Yuri Ivanov
MERL Research
Cambridge, USA
yivanov@merl.com

Louis-Philippe Morency
University of Southern California
Marina del Rey, USA
morency@ict.usc.edu

Anton Nijholt
University of Twente
Enschede, The Netherlands
anijholt@cs.utwente.nl

Maja Pantic
Imperial College
London, United Kingdom
m.pantic@imperial.ac.uk

Ming-Hsuan Yang
Univ. of California
Merced, USA
mhyang@ucmerced.edu

ABSTRACT
Recent efforts in defining ambient intelligence applications based on user-centric concepts, the advent of technology in different sensing modalities as well as the expanding interest in multi-modal information fusion and situation-aware and dynamic vision processing algorithms have created a common motivation across different research disciplines to utilize context as a key enabler of application-oriented vision systems design. Improved robustness, efficient use of sensing and computing resources, dynamic task assignment to different operating modules as well as adaptation to event and user behavior models are among the benefits a vision processing system can gain through the utilization of contextual information. The Workshop on Use of Context in Vision Processing (UCVP) aims to address the opportunities in incorporating contextual information in algorithm design for single or multi-camera vision systems, as well as systems in which vision is complemented with other sensing modalities, such as audio, motion, proximity, occupancy, and others.

Categories and Subject Descriptors
I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding - Motion; I.2.7 [**Artificial Intelligence**]: Natural Language Processing - Discourse; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis; I.5.1 [**Pattern Recognition**]: Models - Statistical; I.5.2 [**Pattern Recognition**]: Pattern Analysis

UCVP '09, November 5, 2009 Boston, MA

General Terms
Algorithms, Human Factors, Theory

Keywords
Contextual information, visual gesture recognition, human-human interaction, context-driven event interpretation, object recognition, statistical relational models, smart homes, intelligent headlight control, camera sensors, machine learning, image/video content analysis, driving context.

1. INTRODUCTION
While offering access to rich information from the environment, computer vision often remains challenged in the face of requirements for accurate and meaningful content interpretation in practical systems. This is partially due to the complexities associated with the choice of early vision features. These features are often subject to becoming unavailable or irrelevant as conditions vary. Another source of challenge is the presence of different possible interpretations from the observations acquired in limited temporal or spatial spans.

On the other hand, three domains of technological advances have culminated in ushering access to information that can enable vision achieve meaningful results in applied situations in novel ways:

- Technologies and methods for multi-modal sensing, multi-camera networks, and data fusion
- Machine learning techniques and multi-layer training methods operating on large data sets
- Information database management tools offering purposeful access to an unprecedented amount of data available as the common knowledge base on the internet

These availabilities combined with the need of many novel applications in ambient intelligence and smart environments to learn and adapt to the behavior models and user preferences create new opportunities for introducing contextual data in vision processing and enabling vision to assume its essential role as an information-rich source for many practical applications.

Context can be defined as prior information extracted from the environment or accumulated in and accessed from a knowledge base. It can also be regarded as logical relationship between the interesting observed entities within a single frame, across a temporal span of frames, or between views in a network of cameras, in each case enabling an efficient and purposeful processing path to the detection or interpretation objective. Contextual information can be exchanged between vision and other sensing modalities such as audio and proximity sensing systems to expedite effective processing. In a broader sense, context can serve both as an output of vision processing setting the real-time situation for high-level reasoning modules, or as a feedback element from higher layers of processing to enable suitable modes of operation within the vision module. User behavior and event models, environmental states and parameters acquired by various sensing methods, logical relationship between objects in physical spaces and in images, consistency between different instances of observation in time and views, and previously interpreted observations are among the multitude forms of context. As such, use of context enables the creation of a broad framework for explicit use of available knowledge in the design of vision systems.

2. UCVP 2009 WORKSHOP

The fields of multi-camera vision, multi-modal systems, smart environments, and human-centric interfaces have seen rapidly increasing interest in recent years as evidenced by the multiple events organized among different research communities. Use of context in vision processing is a subject that finds matching interest in each of the disciplines participating in these research directions. Context-aware processing of visual data allows for complex scene interpretation, event modeling and user behavior routine extraction, as well as efficient selection of vision processing tools and parameters based on the situation, thus making context-based vision processing an attractive approach for a wide number of applications. As such, the UCVP 2009 workshop aims to reach out to the various communities with interest in vision and context, and to act as a medium bridging the gap between these disciplines.

UCVP aims to address the opportunities in incorporating contextual information in algorithm design for single or multi-camera vision systems, as well as systems in which vision is complemented with other sensing modalities, such as audio, motion, proximity, occupancy, and others. The objective of the workshop is to present high-quality contributions describing leading-edge research in the use of context in vision processing.

We distinguished the following topics of interest:

- **Sources of context** (multi-camera networks, multi-modal sensing systems, long-term observation, behavior models, spatial or temporal relationships of objects and events, interaction of user with objects, internet resources as knowledge-base for context extraction)
- **User-centric context** (demographic information, activity, user's emotional state, stated preferences, explicit and implicit interfaces, interaction between users)
- **Uses of context** (context-driven event interpretation, active vision, multi-modal activation, service provision and switching based on context, response and interaction with user, detection of abnormal behavior, active sensing, task assignment to different sensing modules, guided vision based on high-level reasoning, user behavior modeling, applications in smart environments, human-computer interfaces)

The workshop aims to encourage collaboration between researchers in different areas of computer vision and related disciplines. In addition, by introducing topics of emerging applications in smart environments, multi-camera networks, and multi-modal sensing as sources of context in vision, the workshop aims to extend the notion of context-based vision processing to include high-level and application-driven information extraction and fusion.

3. UCVP 2009 PRESENTATIONS

In order to allow intensive discussions among the participants, in addition to two long invited presentations, we only selected five papers for inclusion in the program. These papers can be found in these proceedings: [1], [2], [3], [4]. The invited talks were given by Alex (Sandy) Pentland and Antonio Torralba, both from Massachusetts Institute of Technology in Cambridge, USA. Below the summaries of the talks can be found.

Antonio Torralba. Massachusetts Institute of Technology. Understanding Visual Scenes

Human visual scene understanding is remarkable: with only a brief glance at an image, an abundance of information is available - spatial structure, scene category and the identity of main objects in the scene. In traditional computer vision, scene and object recognition are two visual tasks generally studied separately. However, it is unclear whether it is possible to build robust systems for scene and object recognition, matching human performance, based only on local representations. Another key component of machine vision algorithms is the access to data that describe the content of images. As the field moves into integrated systems that try to recognize many object classes and learn about contextual relationships between objects, the lack of large annotated datasets hinders the fast development of robust solutions. In the early days, the first challenge a computer vision researcher would encounter would be the difficult task of digitizing a photograph. Even once a picture was in digital form, storing a large number of pictures (say six) consumed most of the available computational resources. In addition to the algorithmic advances required to solve object recognition, a key component to progress is access to data in order to train computational models for the different object classes. This situation has dramatically changed in the last decade, especially via the internet, which has given computer vision researchers access to billions of images and videos. In this talk I will describe recent work on visual scene understanding that try to build integrated models for scene and object recognition, emphasizing the power of large database of annotated images in computer vision.

Alex (Sandy) Pentland. Massachusetts Institute of Technology. Honest Signals: Social Context in Visual Processing

Abstract: We have developed the technology of reality mining, which uses sensor data to extract subtle patterns that predict future human behavior. These predictive patterns are based on "honest signals," which are human behaviors that evolved from ancient primate signaling mechanisms, and which are major factors in human decision making in everything from job interviews to first dates. By building interfaces based on honest signals, we have been able to obtain dramatic improvements in human-machine systems.

The organizers of UCVP 2009 are satisfied with the workshop and hope it will be the start of a workshop series on this topic.

4. ACKNOWLEDGMENTS

Our thanks go to the UCVP 2009 Program Committee members: Stan Birchfield (Clemson University, USA), Tanzeem Choudhury (Dartmouth College, USA), Bill Christmas (University of Surrey, UK), Maurice Chu (PARC, Palo Alto, USA), David Demirdjian (MIT, USA), Abhinav Gupta (University of Maryland, USA), Ronald Poppe (TU Delft, The Netherlands), Paolo Remagnino (Kingston University, UK), Neil Robertson (Heriot-Watt University, UK), Michael S. Ryoo (ETRI, Korea), Stan Sclaroff (Boston University, USA), Rainer Stiefelhagen, University of Karlsruhe, Germany), YingLi Tian (CCNY, New York), Fernando de la Torre (CMU, USA).

5. REFERENCES

[1] Kreutzmann, A., Terzic, K., and Neumann, B. 2009 Context-aware Classification for Incremental Scene Interpretation. Proceedings Use of Context in Visual Processing (UCVP 2009). Workshop organized at ICMI-MLMI 2009, Cambridge, Mass., USA, November 5, 2009. ACM Digital Library, H. Aghajan, R. Braspenning, Y. Ivanov, L.-P. Morency, A. Nijholt, M. Pantic, and Ming-Hsuan Yang (Eds.), November 2009.

[2] Li, Y. and Pankanti, S. 2009 Intelligent Headlight Control Using Camera Sensors. Proceedings Use of Context in Visual Processing (UCVP 2009). Workshop organized at ICMI-MLMI 2009, Cambridge, Mass., USA, November 5, 2009. ACM Digital Library, H. Aghajan, R. Braspenning, Y. Ivanov, L.-P. Morency, A. Nijholt, M. Pantic, and Ming-Hsuan Yang (Eds.), November 2009.

[3] Morency, L.-P. 2009 Data-driven Context Representation for Head Gesture Recognition during Multi-Party Interaction. Proceedings Use of Context in Visual Processing (UCVP 2009). Workshop organized at ICMI-MLMI 2009, Cambridge, Mass., USA, November 5, 2009. ACM Digital Library, H. Aghajan, R. Braspenning, Y. Ivanov, L.-P. Morency, A. Nijholt, M. Pantic, and Ming-Hsuan Yang (Eds.), November 2009.

[4] Wu, C. and Aghajan, H. 2009 Using Context with Statistical Relational Models - Object Recognition from Observing User Activity in Home Environment Proceedings Use of Context in Visual Processing (UCVP 2009). Workshop organized at ICMI-MLMI 2009, Cambridge, Mass., USA, November 5, 2009. ACM Digital Library, H. Aghajan, R. Braspenning, Y. Ivanov, L.-P. Morency, A. Nijholt, M. Pantic, and Ming-Hsuan Yang (Eds.), November 2009.