Vision-Based Interaction

Synthesis Lectures on Computer Vision

Synthesis Lectures on Computer Vision is edited by Gérard Medioni of the University of Southern California and Sven Dickinson of the University of Toronto. The series will publish 50- to 150 page publications on topics pertaining to computer vision and pattern recognition. The scope will largely follow the purview of premier computer science conferences, such as ICCV, CVPR, and ECCV. Potential topics include, but not are limited to:

- Applications and Case Studies for Computer Vision
- Color, Illumination, and Texture
- Computational Photography and Video
- Early and Biologically-inspired Vision
- Face and Gesture Analysis
- Illumination and Reflectance Modeling
- Image-Based Modeling
- Image and Video Retrieval
- Medical Image Analysis
- Motion and Tracking
- Object Detection, Recognition, and Categorization
- Segmentation and Grouping
- Sensors
- Shape-from-X
- Stereo and Structure from Motion
- Shape Representation and Matching

- Statistical Methods and Learning
- Performance Evaluation
- Video Analysis and Event Recognition

Vision-Based Interaction
Matthew Turk and Gang Hua
2013

Camera Networks: The Acquisition and Analysis of Videos over Wide Areas
Amit K. Roy-Chowdhury and Bi Song
2012

Deformable Surface 3D Reconstruction from Monocular Images
Mathieu Salzmann and Pascal Fua
2010

Boosting-Based Face Detection and Adaptation
Cha Zhang and Zhengyou Zhang
2010

Image-Based Modeling of Plants and Trees
Sing Bing Kang and Long Quan
2009

Vision-Based Interaction
Matthew Turk and Gang Hua

ISBN: 978-3-031-00684-5 paperback
ISBN: 978-3-031-01812-1 ebook

DOI 10.1007/978-03-031-01812-1

A Publication in the Springer series
SYNTHESIS LECTURES ON COMPUTER VISION

Lecture #5
Series Editors: Gérard Medioni, *University of Southern California*
Sven Dickinson, *University of Toronto*
Series ISSN
Synthesis Lectures on Computer Vision
Print 2153-1056 Electronic 2153-1064

Vision-Based Interaction

Matthew Turk
University of California, Santa Barbara

Gang Hua
Stevens Institute of Technology

SYNTHESIS LECTURES ON COMPUTER VISION #5

ABSTRACT

In its early years, the field of computer vision was largely motivated by researchers seeking computational models of biological vision and solutions to practical problems in manufacturing, defense, and medicine. For the past two decades or so, there has been an increasing interest in computer vision as an input modality in the context of human-computer interaction. Such *vision-based interaction* can endow interactive systems with visual capabilities similar to those important to human-human interaction, in order to perceive non-verbal cues and incorporate this information in applications such as interactive gaming, visualization, art installations, intelligent agent interaction, and various kinds of command and control tasks. Enabling this kind of rich, visual and multimodal interaction requires interactive-time solutions to problems such as detecting and recognizing faces and facial expressions, determining a person's direction of gaze and focus of attention, tracking movement of the body, and recognizing various kinds of gestures.

In building technologies for vision-based interaction, there are choices to be made as to the range of possible sensors employed (e.g., single camera, stereo rig, depth camera), the precision and granularity of the desired outputs, the mobility of the solution, usability issues, etc. Practical considerations dictate that there is not a one-size-fits-all solution to the variety of interaction scenarios; however, there are principles and methodological approaches common to a wide range of problems in the domain. While new sensors such as the Microsoft Kinect are having a major influence on the research and practice of vision-based interaction in various settings, they are just a starting point for continued progress in the area.

In this book, we discuss the landscape of history, opportunities, and challenges in this area of vision-based interaction; we review the state-of-the-art and seminal works in detecting and recognizing the human body and its components; we explore both static and dynamic approaches to "looking at people" vision problems; and we place the computer vision work in the context of other modalities and multimodal applications. Readers should gain a thorough understanding of current and future possibilities of computer vision technologies in the context of human-computer interaction.

KEYWORDS

computer vision, vision-based interaction, perceptual interface, face and gesture recognition, movement analysis

MT: To K, H, M, and L

GH: To Yan and Kayla, and my family

Contents

Preface

Like many areas of computing, vision-based interaction has found motivation and inspiration from authors and filmmakers who have painted compelling pictures of future technology. From *2001: A Space Odyssey* to *The Terminator* to *Minority Report* to *Iron Man*, audiences have seen computers interacting with people visually in natural, human-like ways: recognizing people, understanding their facial expressions, appreciating their artwork, measuring their body size and shape, and responding to gestures. While this often works out badly for the humans in these stories, presumably this is not the fault of the interface, and in many cases these futuristic visions suggest useful and desirable technologies to pursue.

Perusing the proceedings of the top computer vision conferences over the years shows just how much the idea of computers looking at people has influenced the field. In the early 1990s, a relatively small number of papers had images of people in them, while the vast majority had images of generic objects, automobiles, aerial views, buildings, hallways, and laboratories. (Notably, there were many papers back then with no images at all!) In addition, computer vision work was typically only seen in computer vision conferences. Nowadays, conference papers are full of images of people—not all in the context of interaction, but for a wide range of scenarios where people are the main focus of the problems being addressed—and computer vision methods and technologies appear in a variety of other research venues, especially including CHI (human-computer interaction), SIGGRAPH (computer graphics and interactive techniques) and multimedia conferences, as well as conferences devoted exclusively to these and related topics, such as FG (face and gesture recognition) and ICMI (multimodal interaction). It seems reasonable to say that people have become a main focus (if not *the* main focus) of computer vision research and applications.

Part of the reason for this is the significant growth in consumer-oriented computer vision—solutions that provide tools to improve picture taking, organizing personal media, gaming, exercise, etc. Cameras now find faces, wait for the subjects to smile, and do automatic color balancing to make sure the skin looks about right. Services allow users to upload huge amounts of image and video data and then automatically identify friends and family members and link to related stored images and video. Video games now track multiple players and provide live feedback on performance, calorie burn, and such. These consumer-oriented applications of computer vision are just getting started; the field is poised to contribute in many diverse and significant ways in the years to come. An additional benefit for those of us who have been in the field for a while is that we can finally explain to our relatives what we do, without the associated blank stares.

The primary goals of this book are to present a bird's eye view of vision-based interaction, to provide insight into the core problems, opportunities, and challenges, and to supply a snapshot of key methods and references at this particular point in time.

While the machines are still on our side.

Matthew Turk and Gang Hua
September 2013

Acknowledgments

We would firstly like to thank Gerard Medioni and Sven Dickinson, the editors of this Synthesis Lectures on Computer Vision series, for inviting us to contribute to the series. We are grateful to the reviewers, who provided us with constructive feedback that made the book better. We would also like to thank all the people who granted us permission to use their figures in this book. Without their contribution, it would have been much more difficult for us to complete the manuscript. We greatly appreciate the support, patience, and help of our editor, Diane Cerra, at every phase of writing this book. Last but not least, we would like to thank our families for their love and support.

We would like to acknowledge partial support from the National Science Foundation.

Matthew Turk and Gang Hua
September 2013

Figure Credits

Figures 1.2 a, b from *2001: A Space Odyssey*, 1968. Metro-Goldwyn-Mayer Inc., 3 April 1968; LP36136 (in copyright registry) Copyright © Renewed 1996 by Turner Entertainment Company.

Figure 1.2 c from *The Terminator*, 1984. Copyright © 2011 by Annapurna Pictures.

Figure 1.2 d from *Minority Report*, 2002. Copyright © 2002 BY Dreamworks LLC and Twentieth Century Fox Film Corporation.

Figures 1.2 e, f from *Iron Man*, 2008. Copyright © 2008 by Marvel.

Figures 1.3 a, b from Myron Krueger, *Videoplace*, 1970. Used with permission.

Figures 1.4 a, b courtesy of Irfan Essa.

Figures 1.4 c, d courtesy of Jim Davis

Figures 1.4 e, f courtesy of Christopher Wren

Figures 2.2 a, b and 2.3 based on Viola, et al: Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2001*, volume 1, pages 511-518. Copyright © 2001 IEEE. Adapted courtesy of Viola, P. A. and Jones, M. J.

Figures 2.4 a, b, c, d, e, f, g and 2.5 from Hua, et al: A robust elastic and partial matching metric for face recognition. *Proceedings of the IEEE International Conference on Computer Vision, 2009*. Copyright © 2009 IEEE. Used with permission.

Figure 2.12 based on Song, et al: Learning universal multi-view age estimator by video contexts. *Proceedings of the IEEE International Conference on Computer Vision, 2011*. Copyright © 2011 IEEE. Adapted courtesy of Song, Z., Ni, B., Guo, D., Sim, T., and Yan, S.

Figure 2.13 from Jesorsky, et al: Robust face detection using the hausdorff distance. *Audio- and Video-Based Biometric Person Authentication: Proceedings of the Third International Conference, AVBPA 2001 Halmstad, Sweden, June 6–8, 2001*, pages 90-95. Copyright © 2001, Springer-Verlag Berlin Heidelberg. Used with permission. DOI: 10.1007/3-540-45344-X_14

Figure 2.14 based on Chen, J. and Ji, Q. Probabilistic gaze estimation without active personal calibration. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011*. Copyright © 2011 IEEE. Adapted courtesy of Chen, J. and Ji, Q.

Figures 2.15 a, b, c, d, e from Mittal, et al: Hand detection using multiple proposals. *British Machine Vision Conference, 2011*. Copyright and all rights therein are retained by authors. Used courtesy of Mittal, A., Zisserman, A., and Torr, P. H. S. `http://www.robots.ox.ac.uk/~vgg/publications/2011/Mittal11/`

Figure 2.16 Wachs, et al: Vision-based hand-gesture applications. *Communications of the ACM*, 54(2), 60-72. Copyright © 2011, Association for Computing Machinery, Inc. Reprinted by permission. DOI: 10.1145/1897816.1897838

Figure 2.17 from Felzenszwalb, et al: Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), 1627-1645. Copyright © 2010 IEEE. Used with permission. DOI: 10.1109/TPAMI.2009.167

Figure 2.18 from Codasign. *Skeleton tracking with the kinect*. Used with permission. URL: `http://learning.codasign.com/index.php?title=Skeleton_Tracking_with_the_Kinect`

Figure 3.1 from *Kinect Rush: A Disney Pixar Adventure*. Copyright © 2012 Microsoft Studio.

Figure 3.2 from Freeman, et al: Television control by hand gestures. *IEEE International Workshop on Automatic Face and Gesture Recognition, Zurich*. Copyright © 1995 IEEE. Used with permission.

Figures 3.3 a, b from Iannizzotto, et al: A vision-based user interface for real-time controlling toy cars. *10th IEEE Conference on Emerging Technologies and Factory Automation, 2005 (ETFA 2005)*, volume 1. Copyright © 2005 IEEE. Used with permission.

Figure 3.4 from Stenger, et al: A vision-based remote control. In R. Cipolla, S. Battiato, and G. Farinella (Eds.), *Computer Vision: Detection, Recognition and Reconstruction*, pages 233-262. Springer Berlin / Heidelberg. Copyright © 2010, Springer-Verlag Berlin Heidelberg. Used with permission. DOI: 10.1007/978-3-642-12848-69

Figures 3.5 a, b from Tu, et al: Face as mouse through visual face tracking. *Computer Vision and Image Understanding*, 108(1-2), 35-40. Copyright © 2007 Elsevier Inc. Reprinted with permission. DOI: 10.1016/j.cviu.2006.11.007

Figure 3.6 a from Marcel, et al: Hand gesture recognition using input-output hidden markov models. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000*. Copyright © 2000 IEEE. Used with permission. DOI: 10.1109/AFGR.2000.840674

Figure 3.6 b based on Marcel, et al: Hand gesture recognition using input-output hidden markov models. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000*. Copyright © 2000 IEEE. Adapted courtesy of Marcel, S., Bernier, O., and Collobert, D. DOI: 10.1109/AFGR.2000.840674

Figure 3.7 based on Rajko, et al: Real-time gesture recognition with minimal training requirements and on-line learning. *IEEE Conference on Computer Vision and Pattern Recognition, 2007*. Copyright © 2007 IEEE. Adapted courtesy of Rajko, S., Gang Qian, Ingalls, T., and James, J.

Figure 3.8 a based on Elgammal, et al: Learning dynamics for exemplar-based gesture recognition. *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Copyright © 2003 IEEE. Adapted courtesy of Elgammal, A., Shet, V., Yacoob, Y., and Davis, L. S. DOI: 10.1109/CVPR.2003.1211405

Figure 3.8 b from Elgammal, et al: Learning dynamics for exemplar-based gesture recognition. *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Copyright © 2003 IEEE. Used with permission. DOI: 10.1109/CVPR.2003.1211405

Figure 3.9 from Wang, et al: Hidden conditional random fields for gesture recognition. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Copyright © 2006 IEEE. Used with permission. DOI: 10.1109/CVPR.2006.132

Figures 3.10 and 3.11 b based on Shen, et al: (2012). Dynamic hand gesture recognition: An exemplar based approach from motion divergence fields. *Image and Vision Computing: Best of Automatic Face and Gesture Recognition 2011*, 30(3), 227-235. Copyright © 2011 Elsevier B.V. Adapted courtesy of Shen, X., Hua, G., Williams, L., and Wu, Y.

Figures 3.11 a, c from Shen, et al: (2012). Dynamic hand gesture recognition: An exemplar based approach from motion divergence fields. *Image and Vision Computing: Best of Automatic Face and Gesture Recognition 2011*, 30(3), 227-235. Copyright © 2011 Elsevier B.V. Used courtesy of Shen, X., Hua, G., Williams, L., and Wu, Y.

Figure 3.12 based on Hua, et al: Peye: Toward a visual motion based perceptual interface for mobile devices. *Proceedings of the IEEE International Workshop on Human Computer Interaction 2007*, pages 39-48. Copyright © 2007 IEEE. Adapted courtesy of Hua, G., Yang, T.-Y., and Vasireddy, S.

Figures 3.13 a, b from Starner, et al: Real-time American sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12), 1371-1375. Copyright © 1998 IEEE. Used with permission. DOI: 10.1109/34.735811

Figure 3.14 from Vogler et al: A framework for recognizing the simultaneous aspects of American sign language. *Computer Vision and Image Understanding*, 81(3), 358-384. Copyright © 2001 Academic Press. Used with permission.

Figure 3.15 based on Vogler et al: A framework for recognizing the simultaneous aspects of American sign language. *Computer Vision and Image Understanding*, 81(3), 358-384. Copyright © 2001 Academic Press. Adapted courtesy of Vogler, C. and Metaxas, D.

Figure 4.1 from Bolt, R. A. (1980). "Put-that-there": Voice and gesture at the graphics interface. *Proceeding SIGGRAPH '80 Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, pages 262-270. Copyright © 1980, Association for Computing Machinery, Inc. Reprinted by permission. DOI: 10.1145/800250.807503

Figure 4.2 from Sodhi, et al: Aireal: Interactive tactile experiences in free air. *ACM Transactions on Graphics (TOG) - SIGGRAPH 2013 Conference Proceedings*, 32(4), July 2013, Article No. 134. Copyright © 2013, Association for Computing Machinery, Inc. Reprinted by permission. DOI: 10.1145/2461912.2462007

Figure 5.1 a Copyright © 2010 Microsoft Corporation. Used with permission.

Figure 5.1 b courtesy Cynthia Breazeal.

Figure 5.2 d Copyright ©2013 Microsoft Corporation. Used with permission.