Journal of Ambient Intelligence and Smart Environments 3 (2011) 187–191
DOI 10.3233/AIS-2011-0113
IOS Press

Computer vision for ambient intelligence

Albert Ali Salah[a,*], Theo Gevers[b], Nicu Sebe[c] and Alessandro Vinciarelli[d]

[a] *Informatics Institute – ISLA, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands*
[b] *Informatics Institute – ISLA, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands*
E-mail: th.gevers@uva.nl
[c] *Department of Information Engineering and Computer Science, University of Trento, Via Sommarive,*
14 I-38123 Povo, Italy
E-mail: nicu.sebe@disi.unitn.it
[d] *Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, Scotland*
E-mail: vincia@dcs.gla.ac.uk

Abstract. A natural way of conceptualizing ambient intelligence is by picturing an active environment with access to perceptual input, not via eyes and ears, but by their technological counterparts. Computer vision is an essential part of building context-aware environments that adapt and anticipate their human users by understanding their behavior. This thematic issue explores state-of-the-art computer vision approaches for ambient intelligence applications.

Keywords: Computer vision, ambient intelligence, activity analysis, object detection, gesture recognition

1. Introduction

The vision of ambient intelligence (AmI) revolves around understanding human behavior in its many social settings, and aims to provide humans with socially aware cognitive systems, enabling a shift from usability to sociability, or to ambient culture. One can envision seamlessly integrated plug and play devices that can be used to endow a given environment with an awareness of the physical, functional, temporal and social organization of its internal domestic dynamics, as well as the personalities and social relationships of its inhabitants, providing a vast array of new services for ordinary people.

While many novel sensory modalities are offering the designers of AmI applications new possibilities for realizing these goals, computer vision remains the most important modality in providing rich information in an unobtrusive manner. Starting from the 80s, researchers were concerned with the creation of smart environments; Bolt's *Put-that-there* system of 1980 implemented a media environment in which the user could point to an item and use voice commands to affect content changes [6]. At the time the computational resources for a computer vision based solution were too demanding, but now real-time visual tracking of body parts for one or multiple persons is a possibility, and already one sees an explosion in applications that rely on these technologies, as well as specialized hardware (e.g. Microsoft Kinect) that simplify their deployment. This is but one area for which progress in computer vision provides immediately useful tools. Computer vision has applications for ambient assisted living [3], human-computer interaction [10,31], surveillance and identification [33], abnormal event detection [20], and for many more challenges falling under the domain of AmI.

Here we provide a short overview of the most recent research directions that are relevant for ambient intelligence, and summarize the contributions offered in the *Computer Vision for Ambient Intelligence* thematic issue[1].

[*] Corresponding author. E-mail: a.a.salah@uva.nl.

[1] The thematic issue follows the Int. Workshop on Human Behavior Understanding [28], organized as a satellite to ICPR 2010. There was a total of 12 submissions to the thematic issue.

2. Computer vision perspective

It can be argued that progress in basic computer vision tools, like better feature descriptors and improved computer vision algorithms, directly have a bearing on ambient intelligence research relying on such tools. An example is the space-time interest points introduced by Laptev, which proved to be an adequate descriptor to recognize actions from visual input [14].

A second research front is in sensor networks in general and camera networks in particular [23]. Recent work in computer vision goes beyond processing the input of a single camera, and focuses on seamless integration of multiple inputs [4,34,36].

A major concern to using camera networks for ambient intelligence is privacy. However, progress in vision related hardware allows using embedded processors, both reducing the communication load on the network infrastructure, and making pre-processing to address privacy concerns possible [3]. Embedded or GPU-based implementations of feature extraction algorithms can also greatly improve the speed of processing, and enable applications relying on rapid action recognition [26].

3. Human behavior understanding perspective

Human behavior has been a focus of computer vision research for a long time, mostly on a personal signal level, where the face and body of a person are tracked and analysed for a specific purpose. Automatic classification of human behavior involves understanding of bodily motion [19,41], gestures and signs [21], analysis of facial expressions [30] and other affective signals [40]. On a higher level, these signals are integrated with the contextual properties of an application domain, to constrain the otherwise immense variation in expressive human behavior [22].

Research now moves towards analysis in more natural settings with uncontrolled conditions, adding more stringent real-time constraints and most importantly, interaction dynamics. The relative position of people in interaction, their postures, gestures, non-verbal behavior and the way they respond to each other, all carry significant social cues which are essential to correctly infer contextual properties of the interaction. The recently flourishing field of social signal processing is very relevant for AmI researchers, as it attempts to systematically categorize these signals, developing tools for their automatic recognition [29,37].

The emphasis on human behavior is one of the key issues that separates ambient intelligence from ubiquitous and pervasive computing [1]. In order to be context aware, personalized, adaptive and anticipatory, applications of ambient intelligence require a lot of information about the actors in a setting. Computer vision offers a number of solutions in answering some of W5+ questions:

- **Who?** Person identification, face and behavior-based biometrics [11,42]. Visual biometrics can mostly handle the relatively small number of users in a smart environment.
- **What?** Event, action and activity recognition [2, 5,12,18,25,39], object recognition for human-object interactions [13,15]. Progress is also driven by the multimedia retrieval community, which seeks to find ways of mining large video collections for events and activities. For instance, the TRECVID video retrieval evaluation challenge introduced a multimedia event detection task in its benchmarks in 2010[2]. Continuous camera recordings (including wearable cameras that are used for life-logging) can produce enormous amounts of data, making an information retrieval approach essential [8].
- **Where?** Detection and tracking of people [4], estimation of focus of attention [35]. While visual tracking of humans with calibrated cameras is largely solved for constrained settings, tracking multiple persons in complex activities is still challenging. Efforts like the HumanEVA dataset for evaluating human pose and tracking push the field forward (See [32] and references therein, as well as [27] for other benchmarking efforts).
- **How?** Expression, gesture and movement recognition [21,24,40], interaction analysis [9]. Liu et al. recently proposed *action attributes* (e.g. single leg motion, arm over shoulder motion) as a facilitating intermediate representation for action recognition [16]. While low level spatio temporal features and bag-of-words approaches seem to be good in recognizing some actions, higher level representations are required for more complex actions and generalization [17]. Nonverbal interaction analysis concerns interaction management via detection of addressing and turn taking, as well as automatic estimation of internal states (e.g. interest, boredom, agreement), person-

[2] http://www-nlpir.nist.gov/projects/tv2010/tv2010.html

ality traits (e.g. dominance, extroversion, locus-of-control), and social roles [9]. While audio is dominantly used in this kind of research, a great number of facial actions, postures, head and hand actions are also revealing [7].

4. Contributions to the thematic issue

Unconscious or communicative employment of head and hand actions are relevant in human-human communications. Consequently, pre-defined sets of head and hand actions can serve as a natural human-machine interface in a smart environment. Richarz and Fink propose a hidden Markov model framework to combine spatiotemporal gesture trajectories sampled from multiple cameras to classify emblematic gestures in "**Visual recognition of 3D emblematic gestures in an HMM framework**". These emblematic gestures are simple and well-defined signs, easier to segment compared to naturally occuring gestures. The proposed framework is evaluated on nine different gestures (e.g. pointing, waving, come and go gestures, etc.), performed by multiple subjects, and under automatic segmentation assumptions. Furthermore, to assess the bias that might arise from closed-set classification, a garbage model is used, where the non-gesture data from the HumanEVA dataset is used for a reject-class. Apart from the proposed method, an important novelty aspect of this work is that it deals with multi-view hand trajectories.

Establishing the context of a visual scene often relies on the correct detection and identification of persons and objects in the scene. While the users of a smart environment may be few in number, there are many potential objects of interaction, and it becomes a challenge to build detectors for each such object in a generic way. In "**A framework for unsupervised training of object detectors from unlabeled surveillance video**", Celik, Hanjalic and Hendriks consider a static camera based monitoring scenario, where unlabeled surveillance footage is used in an unsupervised fashion to build generic object detectors. The advantage of such a scheme is that a particular setting will have its own set of typical objects (e.g. shopping bags in a shopping mall) that would have different levels of presence and relevance for that setting. On the other hand, lack of direct supervision will add noise to the learning process, and this must be explicitly dealt with.

Ceiling-mounted cameras offer a parsimonious solution to the problem of visual coverage in a scene. The CLEAR evaluation campaign of the European CHIL project [38] previously introduced an extensive dataset acquired with such cameras for smart meeting scenarios. In "**A hybrid probabilistic model for person tracking based on a ceiling-mounted camera**", Yan, Weber and Wermter describe probabilistic models for tracking sitting and moving persons under challenging conditions (e.g. changing illumination, repositioned furniture, distracter persons). The proposed tracking system relies on particle filters, and different visual cues (i.e. motion, shape, shape and color memory) are integrated to drive tracking. The multiple object tracking accuracy (MOTA) measure is used to report results on multiple datasets, with a mean MOTA of around 90 per cent for the considered scenarios.

The ceiling-mounted camera is particularly attractive for ambient assisted living scenarios, where it can replace numerous sensors to facilitate management and maintenance of such systems. In "**Video based technology for Ambient Assisted Living: A review of the literature**", Cardinaux, Bhowmik, Abhayaratne, and Hawley review the literature on video-based ambient assisted living (AAL) applications, including multimodal approaches. Monitoring human behavior in an AAL application is not restricted to activities of daily living (ADL), but includes detection of falls, abnormal events, elopement (a concern for elderly with memory issues), as well as physiological monitoring. Cardinaux et al. also discuss privacy and acceptability issues, which is a key element of AAL applications.

5. Conclusions

Computer vision is a mature but vibrant research field. The papers in this thematic issue demonstrate the use of computer vision methods in unobtrusive monitoring of human behavior for various ambient intelligence scenarios. Apart from acceptance and privacy, the main challenge is the scale of vision based solutions, where one can take into account multiple cameras, multiple persons, many objects of interaction and many activities. Obviously, researchers of ambient intelligence will not only profit from domain-specific investigations, but also from core developments of computer vision in the years to come.

References

[1] E. Aarts and B. de Ruyter, New research perspectives on ambient intelligence, *Journal of Ambient Intelligence and Smart Environments* **1**(1) (2009), 5–14.

[2] J.K. Aggarwal and M.S. Ryoo, Human activity analysis: A review, *ACM Computing Surveys* **43**(3) (2011), 16.

[3] H. Aghajan, J.C. Augusto, C. Wu, P. McCullagh, and J.A. Walkden, Distributed vision-based accident management for assisted living, in: *Proc. of the 5th International Conference on Smart Homes and Health Telematics*, Springer-Verlag, 2007, pp. 196–205.

[4] H.K. Aghajan and A. Cavallaro, *Multi-Camera Networks: Principles and Applications*, Academic Press, 2009.

[5] S. Ali and M. Shah, Human action recognition in videos using kinematic features and multiple instance learning, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **32**(2) (2008), 288–303.

[6] R.A. Bolt, "Put-that-there": Voice and gesture at the graphics interface, in: *Proc. 7th Annual Conf. on Computer Graphics and Interactive Techniques*, 1980, pp. 262–270.

[7] K. Bousmalis, M. Mehu, and M. Pantic, Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools, in: *Proc. 3rd Inf. Conf. on Affective Computing and Intelligent Interaction and Workshops*, IEEE, 2009.

[8] D. Byrne, A.R. Doherty, C.G.M. Snoek, G.J.F. Jones, and A.F. Smeaton, Everyday concept detection in visual lifelogs: Validation, relationships and trends, *Multimedia Tools and Applications* **49**(1) (2010), 119–144.

[9] D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: A review, *Image and Vision Computing* **27**(12) (2009), pp. 1775–1787.

[10] A. Jaimes and N. Sebe, Multimodal human-computer interaction: A survey, *Computer Vision and Image Understanding* **108**(1–2) (2007), 116–134.

[11] A.K. Jain, P.J. Flynn, and A.A. Ross, *Handbook of Biometrics*, Springer-Verlag, New York, Inc., 2008.

[12] I.N. Junejo, E. Dexter, I. Laptev, and P. Pérez, View-independent action recognition from temporal self-similarities, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **33**(1) (2011), 172–185.

[13] H. Kjellström, D. Kragić, and M.J. Black, Tracking people interacting with objects, in: *IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[14] I. Laptev, On space-time interest points, *International Journal of Computer Vision* **64**(2) (2005), 107–123.

[15] L.J. Li, R. Socher, and L. Fei-Fei, Towards total scene understanding: Classification, annotation and segmentation in an automatic framework, in: *IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 2036–2043.

[16] J. Liu, B. Kuipers, and S. Savarese, Recognizing human actions by attributes, in: *IEEE Conf. Computer Vision and Pattern Recognition*, 2011.

[17] J. Liu, M. Shah, B. Kuipers, and S. Savarese, Cross-view action recognition via view knowledge transfer, in: *IEEE Conf. Computer Vision and Pattern Recognition*, 2011.

[18] M. Marszałek, I. Laptev, and C. Schmid, Actions in context, in: *IEEE Conf. Computer Vision & Pattern Recognition*, 2009.

[19] T.B. Moeslund and E. Granum, A survey of computer vision-based human motion capture, *Computer Vision and Image Understanding* **81**(3) (2001), 231–268.

[20] F. Nater, H. Grabner, and L. Van Gool, Exploiting simple hierarchies for unsupervised human behavior analysis, in: *IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[21] S. O'Hara, Y.M. Lui, and B.A. Draper, Unsupervised learning of human expressions, gestures, and actions, in: *IEEE Int. Conf. on Automatic Face & Gesture Recognition and Workshops*, IEEE, 2011.

[22] M. Pantic, A. Pentland, A. Nijholt, and T. Huang, Human computing and machine understanding of human behavior: A survey, *Artifical Intelligence for Human Computing*, 2007, pp. 47–71.

[23] E.J. Pauwels, A.A. Salah, and R. Tavenard, Sensor networks for ambient intelligence, in: *IEEE 9th Workshop on Multimedia Signal Processing*, 2007, pp. 13–16.

[24] V. Pavlovic, R. Sharma, and T.S. Huang, Visual interpretation of hand gestures for human-computer interaction: A review, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **19**(7) (1997), 677–695.

[25] R. Poppe, A survey on vision-based human action recognition, *Image and Vision Computing* **28**(6) (2010), 976–990.

[26] M. Rofouei, M. Moazeni, and M. Sarrafzadeh, Fast GPU-based space-time correlation for activity recognition in video sequences, in: *IEEE/ACM/IFIP Workshop on Embedded Systems for Real-Time Multimedia*, 2008, pp. 33–38.

[27] A.A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli, Challenges of human behavior understanding, in: *Proc. Int. Workshop on Human Behavior Understanding*, LNCS, Vol. 6219, Springer, 2010, pp. 1–12.

[28] A.A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli, *Human Behavior Understanding: First International Workshop, HBU 2010*, LNCS, Vol. 6219, Springer-Verlag, New York, 2010.

[29] A.A. Salah, M. Pantic, and A. Vinciarelli, Recent developments in social signal processing, in: *IEEE Int. Conf. on Systems, Man, and Cybernetics*, IEEE, 2011.

[30] A.A. Salah, N. Sebe, and T. Gevers, Communication and automatic interpretation of affect from facial expressions, in: *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives*, D. Gökçay and G. Yıldırım, eds, IGI Global, 2010.

[31] N. Sebe, Multimodal interfaces: Challenges and perspectives, *Journal of Ambient Intelligence and Smart Environments* **1**(1) (2009), 23–30.

[32] L. Sigal and M.J. Black, Guest editorial: state of the art in image-and video-based human pose and motion estimation, *International Journal of Computer Vision* **87**(1) (2010), 1–3.

[33] L. Snidaro, C. Micheloni, and C. Chiavedale, Video security for ambient intelligence, *IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans* **35**(1) (2005), 133–144.

[34] S. Soro and W. Heinzelman, A survey of visual sensor networks, *Advances in Multimedia* (2009), 1–21.

[35] R. Stiefelhagen, J. Yang, and A. Waibel, Estimating focus of attention based on gaze and sound, in: *Workshop on Perceptive User Interfaces (PUI '01)*, 2001.

[36] M.M. Trivedi, K.S. Huang, and I. Mikic, Dynamic context capture and distributed video arrays for intelligent spaces, *IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans* **35**(1) (2005), 145–163.

[37] A. Vinciarelli, M. Pantic, and H. Bourlard, Social signal processing: Survey of an emerging domain, *Image and Vision Computing* **27**(12) (2009), 1743–1759.

[38] A. Waibel and R. Stiefelhagen, *Computers in the Human Interaction Loop*, Springer, 2009.

[39] D. Weinland, R. Ronfard, and E. Boyer, Free viewpoint action recognition using motion history volumes, *Computer Vision and Image Understanding* **104**(2–3) (2006), 249–257.

[40] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **31**(1) (2009), 39–58.

[41] T. Zhao and R. Nevatia, Tracking multiple humans in complex situations, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **26**(9) (2004), 1208–1221.

[42] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld, Face recognition: A literature survey, *ACM Computing Surveys* **35**(4) (2003), 399–458.