ICIP 2019 Turorial: Multisensory Video Processing and Learning for Human-Robot Interaction

List of References

Tutorial Slides: http://cvsp.cs.ntua.gr/icip2019

Petros Maragos and Petros Koutras

Sunday, September 22, 2019, 14:00 - 17:30

Part 1: Spatio-Temporal Visual Processing

[1] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Amer. A*, 2(2):284–299, 1985.

[2] R. Arandjelovic and A. Zisserman. All about VLAD. In *Proc. CVPR*, 2013.

[3] V. Bettadapura, G. Schindler, T. Plötz, and I. Essa. Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition. In *Proc. IEEE Conf. CVPR*, 2013.

[4] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. In *Proc. IEEE Conf. CVPR*, 2009.

[5] Z. Cai, L. Wang, X. Peng, and Y. Qiao. Multi-view super vector for action recognition. In *Proc. CVPR*, 2014.

[6] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Q.-Q. Chen and Y.-J. Zhang. Sequential segment networks for action recognition. *IEEE Signal Processing Letters*, 24/5:712–716, May 2017.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. CVPR*, 2005.

[9] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[10] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Action spotting and recognition based on a spatiotemporal orientation analysis. *IEEE Trans. PAMI*, 35(3):527–540, 2013.

[11] A. Diba, V. Sharma, and L. Van Gool. Deep temporal linear encoding networks. In *Proc. CVPR*, 2017.

[12] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal Residual Networks for Video Action Recognition. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2016.

[14] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE Conf. Computer Vision & Pattern Recognition*, 2003.

[17] C. Georgakis, P. Maragos, G. Evangelopoulos, and D. Dimitriadis. Dominant spatio-temporal modulations and energy tracking in videos: Application to interest point detection for action recognition. In *Proc. Int'l Conf. Image Processing*, 2012.

[18] T. Hao, D. Wu, Q. Wang, and J. Sun. Multi-view representation learning for multi-view action recognition. *Journal of Visual Communication and Image Representation*, 48:453–460, 2017.

[19] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proc. CVPR*, 2015.

[21] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Proc. CVPR*, 2010.

[22] S. Ji, W. Xu, M. Yang, and K. Yu. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 35:221–231, Jan 2013.

[23] N. Kardaris, V. Pitsikalis, E. Mavroudi, and P. Maragos. Introducing temporal order of dominant visual word sub-sequences for human action recognition. In *Proc. ICIP*, 2016.

[24] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[25] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D-Gradients. In *Proc. BMVC*, 2008.

[26] I. Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] Y. Kong, Z. Ding, J. Li, and Y. Fu. Deeply learned view-invariant features for cross-view action recognition. *IEEE Trans. Image Processing*, 26(6):3028–3037, 2017.

[28] P. Koutras and P. Maragos. A perceptually based spatio-temporal computational framework for visual saliency estimation. *Signal Processing: Image Communication*, 38:15–31, 2015.

[29] P. Koutras and P. Maragos. Susinet: See, understand and summarize it. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[30] I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. ICCV*, 2003.

[31] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. IEEE Conf. CVPR*, 2008.

[32] S.-H. Lee, J.-H. Kim, K. P. Choi, J.-Y. Sim, and C.-S. Kim. Video saliency detection based on spatiotemporal feature learning. In *Proc. Int'l Conf. Image Processing*, 2014.

[33] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. Int'l Conf. on Computer Vision*, 1999.

[34] K. Maninis, P. Koutras, and P. Maragos. Advances on action recognition in videos using and interest point detector based on multiband spatio-temporal energies. In *Proc. Int'l Conf. Image Processing*, 2014.

[35] W. Nie, A. Liu, W. Li, and Y. Su. Cross-view action recognition by cross-domain learning. *Image and Vision Computing*, 55:109–118, 2016.

[36] X. Peng and C. Schmid. Encoding feature maps of CNNs for action recognition. In *Proc. CVPR, THUMOS Challenge 2015 Workshop*, 2015.

[37] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150:109–125, 2016.

[38] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proc. ECCV*, 2010.

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

[40] H. Rahmani and A. Mian. Learning a non-linear knowledge transfer model for cross-view action recognition. In *Proc. CVPR*, 2015.

[41] J. Ray, H. Wang, D. Tran, Y. Wang, M. Feiszli, L. Torresani, and M. Paluri. Scenes-objects-actions: A multi-task, multi-label video dataset. In *Proc. European Conf. on Computer Vision (ECCV)*, 2018.

[42] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *Proc. IEEE Conf. CVPR*, 2012.

[43] A. Sargano, P. Angelov, and Z. Habib. Human action recognition from multiple views based on view-invariant feature descriptor using support vector machines. *Applied Sciences*, 6(10):309, 2016.

[44] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *Proc. ICPR*, 2004.

[45] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proc. Int'l Conf. on Multimedia*, 2007.

[46] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[47] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2015.

[48] C. Theriault, N. Thome, and M. Cord. Dynamic scene classification: Learning motion descriptors with slow features analysis. In *Proc. IEEE Conf. CVPR*, 2013.

[49] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2015.

[50] D. Tran, J. Ray, Z. Shou, S. F. Chang, and M. Paluri. Convnet architecture search for spatiotemporal feature learning. *arXiv preprint arXiv:1708.05038*, 2017.

[51] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[52] G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 40(6):1510–1517, 2018.

[53] H. Wang, A. Klaser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *Proc. CVPR*, 2011.

[54] H. Wang, A. Kläser, C. Schmid, and C. Liu. Dense trajectories and motion boundary descriptors for action recognition. *Int'l J. Comp. Vision*, 103(1):60–79, 2013.

[55] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proc. ICCV*, pages 3551–3558, Dec. 2013.

[56] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. BMVC*, 2009.

[57] L. Wang, W. Li, W. Li, and L. Van Gool. Appearance-and-relation networks for video classification. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[58] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proc. CVPR*, 2015.

[59] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proc. European Conf. on Computer Vision (ECCV)*, 2016.

[60] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *Proc. ICPR Workshop on Learning for Adaptable Visual Systems*, 2004.

[61] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative CNN video representation for event detection. In *Proc. CVPR*, 2015.

[62] Y. Yang, I. Saleemi, and M. Shah. Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. *IEEE Trans. PAMI*, 35(7):1635–1648, 2013.

[63] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

[64] Z. Zhang and D. Tao. Slow feature analysis for human action recognition. *IEEE Trans. PAMI*, 34(3):436–450, 2012.

Part 2: Audio-Visual Processing, Fusion and Perception

[1] P. Aleksic and A. Katsaggelos. Audio-visual biometrics. *Proc. IEEE*, 11:2025–2044, 2006.

[2] Z. Barzelay and Y. Y. Schechner. Harmony in motion. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[3] P. W. Battaglia, R. A. Jacobs, and R. N. Aslin. Bayesian integration of visual and auditory signals for spatial localization. 20(7):1391–1397, July 2003.

[4] M. J. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 25:828–836, 2003.

[5] G. Bouritsas, P. Koutras, A. Zlatintsi, and P. Maragos. Multimodal visual concept learning with weakly supervised techniques. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] J. J. Clark and A. L. Yuille. *Data Fusion for Sensory Information Processing*. Kluwer Academic Publ., 1990.

[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models – their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[8] T. Darrell, J. Fisher, P. Viola, and B. Freeman. Audio-visual segmentation and the cocktail party effect. In *Proc. Int. Conf. on Multimodal Interfaces*, 2000.

[9] J. Driver. Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, 381:66–68, May 1996.

[10] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 445–452. ACM, 2013.

[11] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013.

[12] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.

[13] P. P. Filntisis, A. Katsamanis, and P. Maragos. Photorealistic adaptation and interpolation of facial expressions using hmms and aams for audio-visual speech synthesis. In *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, pages 2941–2945. IEEE, 2017.

[14] P. P. Filntisis, A. Katsamanis, P. Tsiakoulis, and P. Maragos. Video-realistic expressive audio-visual speech synthesis for the greek language. *Speech Communication*, 95:137–152.

[15] C. Garoufis, A. Zlatintsi, K. Kritsis, P. Filntisis, V. Katsouros, , and P. Maragos. An environment for gestural interaction with 3d virual musical instruments as an educational tool. In *Proc. 27th European Conf.(EUSIPCO-19)*.

[16] E. B. Goldstein. *Sensation and Perception*. Wadsworth Publ. Co., California, 1984.

[17] J. Hershey and J. Movellan. Audio-vision: Using audio-visual synchrony to locate sounds. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 1999.

[18] J. M. Hillis, M. O. Ernst, M. S. Banks, and M. S. Landy. Combining sensory information: Mandatory fusion within, but not between, senses. *Science*, 298:1627–1630, 2002.

[19] A. K. Katsaggelos, S. Bahaadini, and R. Molina. Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE*, 103(9):1635–1653, 2015.

[20] A. Katsamanis, G. Papandreou, and P. Maragos. Face active appearance modeling and speech acoustic information to recover articulation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):411–422, 2009.

[21] D. Kersten, P. Mamassian, and A. Yuille. Object perception as bayesian inference. *Annu. Rev. Psychol.*, 55:271–304, 2004.

[22] E. Kidron, Y. Y. Schechner, and M. Elad. Cross-modal localization via sparsity. *IEEE Trans. Signal Process.*, 55(4):1390–1404, Apr. 2007.

[23] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 20(3):226–239, Mar. 1998.

[24] D. C. Knill, D. Kersten, and A. L. Yuille. *Perception as Bayesian Inference*, chapter Introduction: A Bayesian Formulation of Visual Perception, pages 1–21. Cambridge Univ. Press, 1996.

[25] D. C. Knill and W. Richards, editors. *Perception as Bayesian Inference*. Cambridge Univ. Press, 1996.

[26] K. Koffka. *Principles of Gestalt Psychology*. Routledge, 1935, 1999.

[27] W. Köhler. *Gestalt Psychology*. Liveright Publish. Corp., New York, 1947, 1970.

[28] D. Lahat, T. Adali, and C. Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[29] M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Research*, 35(3):389–412, 1995.

[30] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Trans. Multimedia*, 7:907–919, 2005.

[31] P. Maragos, P. Gros, A. Katsamanis, and G. Papandreou. Cross-modal integration for performance improving in multimedia: a review. In *Multimodal processing and interaction*, pages 1–46. Springer, 2008.

[32] P. Maragos, A. Potamianos, and P. Gros. *Multimodal Processing and Interaction: Audio, Video, Text*. Springer-Verlag, New York, 2008.

[33] D. Massaro and D. Stork. Speech recognition and sensory integration. *American Scientist*, 86(3):236–244, 1998.

[34] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 27:305–317, 2005.

[35] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[36] G. Monaci, O. Escoda, and P. Vandergheynst. Analysis of multimodal sequences using geometric video representations. *Signal Processing*, 86:3534–3548, 2006.

[37] G. Monaci and P. Vandergheynst. Audiovisual Gestalts. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW)*, page 200, New York, NY, 2006. IEEE Computer Society.

[38] D. Mumford. *Perception as Bayesian Inference*, chapter Pattern Theory: A unifying perspective, pages 25–61. Cambridge Univ. Press, 1996.

[39] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):423–435, 2009.

[40] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. IEEE*, 92:495–513, 2004.

[41] G. Potamianos, E. Marcheret, Y. Mroueh, V. Goel, A. Koumbaroulis, A. Vartholomaios, and S. Thermos. Audio and visual modality combination in speech processing applications. In *S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Kruger, eds., The Handbook of Multimodal-Multisensor Interfaces, Vol. 1: Foundations, User Modeling, and Multimodal Combinations*. Morgan Claypool Publ., San Rafael, CA, 2017.

[42] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

[43] J. Reynolds, J. Zacks, and T. Braver. A computational model of event segmentation from perceptual prediction. *Cognitive Science*, 31(4):613–643, 2007.

[44] A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics*. Springer-Verlag, 2006.

[45] M. Sargin, Y. Yemez, E. Erzin, and A. Tekalp. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Trans. Multimedia*, 9(7):1396–1403, Nov. 2007.

[46] M. Slaney and M. Covell. FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2001.

[47] C. G. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, January 2005.

[48] R. J. Sternberg. *Cognitive Psychology*. Thomson Wadsworth, 4 edition, 2006.

[49] A. Tsiami, P. Koutras, A. Katsamanis, A. Vatakis, and P. Maragos. A behaviorally inspired fusion approach for computational audiovisual saliency modeling. *Signal Processing: Image Communication*, 76:186 – 200.

[50] E. Tsilionis and A. Vatakis. Multisensory binding: is the contribution of synchrony and semantic congruency obligatory? *Current Opinion in Behavioral Sciences*, 8:7–13, 2016.

[51] A. Vatakis, P. Maragos, I. Rodomagoulakis, and C. Spence. Assessing the effect of physical differences in the articulation of consonants and vowels on audiovisual temporal perception. *J Speech Lang Hear Res*, 2012.

[52] A. Vatakis and C. Spence. Audiovisual synchrony perception for music, speech, and object actions. *Brain research*, 1111(1):134–142, 2006.

[53] A. Vatakis and C. Spence. Crossmodal binding: Evaluating the ?unity assumption? using audiovisual speech stimuli. *Attention, Perception, & Psychophysics*, 69(5):744–756, 2007.

[54] M. T. Wallace, G. E. Roberson, W. D. Hairston, B. E. Stein, J. W. Vaughan, and J. A. Schirillo. Unifying multisensory signals across time and space. *Exp. Brain Research*, 158:252–258, 2004.

[55] J. Wu, J. Cheng, et al. Bayesian co-boosting for multi-modal gesture recognition. *Journal of Machine Learning Research*, 15(1):3013–3036, 2014.

[56] A. L. Yuille and H. H. Bülthoff. *Perception as Bayesian Inference*, chapter Bayesian Decision Theory and Psychophysics, pages 123–161. Cambridge University Press, 1996.

[57] J. M. Zacks, T. S. Braver, M. A. Sheridan, D. I. Donaldson, A. Z. Snyder, J. M. Ollinger, R. L. Buckner, and M. E. Raichle. Human brain activity time-locked to perceptual event boundaries. 4(6):651–655, June 2001.

[58] A. Zlatintsi, P. Filntisis, C. Garoufis, A. Tsiami, K. Kritsis, M. Kaliakatsos-Papakostas, A. Gkiokas, V. Katsouros, and P. Maragos. A web-based real-time kinect application for gestural interaction with virtual musical instruments. In *Proc. of Audio Mostly Conference (AM'18)*.

Part 3 and 4: Audio-Visual HRI: Methodology and Applications in Assistive Robotics

[1] J. Broekens, M. Heerink, , and H. Rosendal. Assistive social robots in elderly care: A review. *Gerontechnology*, 8(2):203–275, 2009.

[2] G. Chalvatzaki, P. Koutras, J. Hadfield, X. S. Papageorgiou, C. S. Tzafestas, and P. Maragos. Lstm-based network for human gait stability prediction in an intelligent robotic rollator. In *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*, pages 4225–4232, 2019.

[3] G. Chalvatzaki, X. Papageorgiou, C. Tzafestas, and P. Maragos. Augmented human state estimation using interacting multiple model particle filters with probabilistic data association. In *Proc. IEEE Int'l Conf. on Robotics & Automation (ICRA-18)*, Brisbane, Australia, 2018.

[4] G. Chalvatzaki, G. Pavlakos, K. Maninis, X. Papageorgiou, V. Pitsikalis, C. Tzafestas, and P. Maragos. Towards an intelligent robotic walker for assisted living using multimodal sensorial data. In *Proc. Int'l Conf. on Wireless Mobile Communication and Healthcare (Mobihealth-14)*, pages 156–159. IEEE, 2014.

[5] A. Dometios, X. Papageorgiou, A. Arvanitakis, C. Tzafestas, and P. Maragos. Real-time end-effector motion behavior planning approach using on-line point-cloud data towards a user adaptive assistive bath robot. In *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS-2017)*, pages 5031–5036. IEEE, 2017.

[6] A. Dometios, A. Tsiami, A. Arvanitakis, P. Giannoulis, X. Papageorgiou, C. Tzafestas, and P. Maragos. Integrated speech-based perception system for user adaptive robot motion

planning in assistive bath scenarios. In *Proc. of the 25th European Signal Proc. Conf. - Workshop: "MultiLearn 2017 - Multimodal processing, modeling and learning for human-computer/robot interaction applications"*, Kos, Greece, Aug.-Sep. 2017.

[7] E. Efthimiou, S.-E. Fotinea, T. Goulas, A.-L. Dimou, M. Koutsombogera, V. Pitsikalis, P. Maragos, and C. Tzafestas. The mobot platform–showcasing multimodality in human-assistive robot interaction. In *Proc. Int'l Conf. on Universal Access in Human-Computer Interaction*, pages 382–391. Springer, 2016.

[8] M. A. Goodrich and A. C. Schultz. Human-robot interaction: A survey. *Found. trends human-computer Interact.*, 1(3):203–275, 2007.

[9] A. Guler, N. Kardaris, S. Chandra, V. Pitsikalis, C. Werner, K. Hauer, C. Tzafestas, P. Maragos, and I. Kokkinos. Human joint angle estimation and gesture recognition for assistive robotic vision. In *Proc. European Conference on Computer Vision*, pages 415–431. Springer, 2016.

[10] R. Kachouie, S. Sedighadeli, R. Khosla, and M.-T. Chu. Socially assistive robots in elderly care: A mixed-method systematic literature review. *Int'l Jour. Human-Computer Interaction*, 30(5):369—-393, 2014.

[11] N. Kardaris, V. Pitsikalis, E. Mavroudi, and P. Maragos. Introducing temporal order of dominant visual word sub-sequences for human action recognition. In *Proc. Int'l Conf. on Image Processing (ICIP-2016)*, pages 3061–3065. IEEE, 2016.

[12] N. Kardaris, I. Rodomagoulakis, V. Pitsikalis, A. Arvanitakis, and P. Maragos. A platform for building new human-computer interface systems that support online automatic recognition of audio-gestural commands. In *Proc. of the 2016 ACM on Multimedia Conf.*, pages 1169–1173. ACM, 2016.

[13] A. Katsamanis, V. Pitsikalis, S. Theodorakis, and P. Maragos. Multimodal gesture recognition. In *The Handbook of Multimodal-Multisensor Interfaces*, pages 449–487. Association for Computing Machinery and Morgan & Claypool, 2017.

[14] A. Kotteritzsch and B. Weyers. Assistive technologies for older adults in urban areas: A literature review. *Cognitive Computation*, 8:299–317, 2016.

[15] P. Maragos, V. Pitsikalis, A. Katsamanis, N. Kardaris, E. Mavroudi, I. Rodomagoulakis, and A. T. 2015. Multimodal sensory processing for human action recognition in mobility assistive robotics. In *Proc. IROS-2015 Workshop on Cognitive Mobility Assistance Robots*, Hamburg, Germany, Sep. 2015.

[16] E. Mordoch, A. Osterreicher, L. Guse, K. Roger, and G. Thompson. Use of social commitment robots in the care of elderly people with dementia: A literature review. *Maturitas*, 74:14–20, 2013.

[17] V. Pitsikalis, A. Katsamanis, S. Theodorakis, and P. Maragos. Multimodal gesture recognition via multiple hypotheses rescoring. *The Journal of Machine Learning Research*, 16(1):255–284, 2015.

[18] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, A. Arvanitakis, and P. Maragos. A multimedia gesture dataset for human robot communication: Acquisition, tools and recognition results. In *Proc. Int'l Conf. on Image Processing (ICIP-2016)*, pages 3066–3070. IEEE, 2016.

[19] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos. Multimodal human action recognition in assistive human-robot interaction. In *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP-16)*, pages 2702–2706. IEEE, 2016.

[20] I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, A. Tsiami, and P. Maragos. Room-localized spoken command recognition in multi-room, multi-microphone environments. *Computer Speech & Language*, 46:419–443, 2017.

[21] F. Rudzicz, R. Wang, M. Begum, , and A. Mihailidis. Speech interaction with personal assistive robots supporting aging at home for individuals with alzheimer's disease. *ACM Trans. Access. Comput.*, 7(2):1–222, 2015.

[22] A. Zlatintsi, I. Rodomagoulakis, P. Koutras, A. C. Dometios, V. Pitsikalis, C. S. Tzafestas, and P. Maragos. Multimodal signal processing and learning aspects of human-robot interaction for an assistive bathing robot. In *Proc. 43rd IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP-18)*, Calgary, Canada, Apr. 2018.

[23] A. Zlatintsi, I. Rodomagoulakis, V. Pitsikalis, P. Koutras, N. Kardaris, X. Papageorgiou, C. Tzafestas, and P. Maragos. Social human-robot interaction for the elderly: two real-life use cases. In *Proc. of the 2017 ACM/IEEE Int'l Conf. on Human-Robot Interaction*, pages 335–336. ACM, 2017.

Audio-Visual HRI in Social Robotics for Child-Robot Interaction

[1] T. Belpaeme, P. Baxter, R. Read, R. Wood, and et al. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, 1(2):33–53, 2012.

[2] N. Efthymiou, P. Koutras, P. P. Filntisis, G. Potamianos, and P. Maragos. Multi-view fusion for action recognition in child-robot interaction. In *Proc. Int'l Conf. on Image Processing (ICIP-18)*, Athens, Greece, Oct. 2018.

[3] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos. Fusing body posture with facial expressions for joint recognition of affect in child-robot interaction. *IEEE Robotics and Automation Letters (RA-L)*, 2019.

[4] P. Giannoulis, G. Potamianos, and P. Maragos. On the joint use of nmf and classification for overlapping acoustic event detection. In *Multidisciplinary Digital Publishing Institute Proceedings*, volume 2, page 90, 2018.

[5] J. Hadfield, G. Chalvatzaki, P. Koutras, M. Khamassi, C. S. Tzafestas, and P. Maragos. A deep learning approach for multi-view engagement estimation of children in a child-robot joint attention task. In *Proc. of 2019 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS-2019)*, 2019.

[6] J. Hadfield, P. Koutras, N. Efthymiou, G. Potamianos, C. Tzafestas, and P. Maragos. Object assembly guidance in child-robot interaction using rgb-d based 3d tracking. In *Proc. of 2018 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS-2018)*, Madrid, Spain, Oct. 2018.

[7] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme. Child speech recognition in human-robot interaction: evaluations and recommendations. In *Proc. on Human Robot Interaction (HRI-17)*, 2017.

[8] A. Potamianos, C. Tzafestas, E. Iosif, F. Kirstein, P. Maragos, K. Dauthenhahn, J. Gustafson, J. Ostergaard, S. Kopp, P. Wik, et al. Babyrobot—next generation social robots: Enhancing communication and collaboration development of td and asd children by developing and commercially exploiting the next generation of human-robot interaction technologies. In *Proc. of the Workshop on Evaluating Child-Robot Interaction (CRI) at the ACM/IEEE Int'l Conf. on Human-Robot Interaction (HRI)*, volume 495, 2016.

[9] B. Robins, K. Dautenhahn, R. T. Boekhorst, and A. Billard. Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Universal Access in the Information Society*, 4(2):105—-120, 2005.

[10] A. Tsiami, P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos. Far-field audio-visual scene perception of multi-party human-robot interaction for children and adults. In *Proc. 43rd IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP-18)*, Calgary, Canada, Apr. 2018.

[11] A. Tsiami, P. Koutras, N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos. Multi3: Multi-sensory perception system for multi-modal child interaction with multiple robots. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA-18)*, Brisbane, Australia, May 2018.