References

[1] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*. PhD thesis, Technische Universität München, 2008.

[2] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3642–3649, IEEE, 2012.

[3] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 833–840, 2011.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing systems*, pp. 1097–1105, 2012.

[5] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (Colorado Springs, CO), pp. 3169–3176, IEEE, Jun 2011.

[6] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of IEEE International Conference on Computer Vision*, (Sydney, AUS), pp. 3551–3558, IEEE, Nov 2013.

[7] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 568–576, 2014.

[8] S. Ramasinghe and R. Rodrigo, "Action recognition by single stream convolutional neural networks: An approach using combined motion and static information," in *Proceedings of the Asian Conference on Pattern Recognition*, pp. 101–105, Nov 2015.

[9] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," in *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*, pp. 90–102, IEEE, 1997.

[10] D. M. Gavrila, "The visual analysis of human movement: A survey," *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[11] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern recognition*, vol. 36, no. 3, pp. 585–601, 2003.

[12] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[13] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, no. 1, pp. 116–134, 2007.

[14] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[15] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[16] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, p. 16, 2011.

[17] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1995–2006, 2013.

[18] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pp. 149–187, Springer, 2013.

[19] J. K. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.

[20] G. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognition*, vol. 47, no. 10, pp. 3343–3361, 2014.

[21] L. Onofri, P. Soda, M. Pechenizkiy, and G. Iannello, "A survey on using domain and contextual knowledge for human activity recognition in video streams," *Expert Systems with Applications*, vol. 63, pp. 97–111, 2016.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 886–893, IEEE, 2005.

[23] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer vision–ECCV 2006*, pp. 404–417, 2006.

[24] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proceeding of the European conference on computer vision*, pp. 428–441, Springer, 2006.

[25] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *Proceeding of the British Machine Vision Conference*, pp. 275–1, British Machine Vision Association, 2008.

[26] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, 2007.

[27] A. Iosifidis, A. Tefas, and I. Pitas, "Activity-based person identification using fuzzy representation and discriminant learning," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 530–542, 2012.

[28] Z. F. Huang, *Latent boosting for action recognition*. PhD thesis, Applied Science: School of Computing Science, 2012.

[29] S. Yi, H. Krim, and L. K. Norris, "Human activity as a manifold-valued random process," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3416–3428, 2012.

[30] S. Wang, Z. Ma, Y. Yang, X. Li, C. Pang, and A. G. Hauptmann, "Semi-supervised multiple feature analysis for action recognition," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 289–298, 2014.

[31] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 3273–3280, IEEE, 2011.

[32] C. Sun and R. Nevatia, "Active: Activity concept transitions in video event classification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 913–920, 2013.

[33] B. Ni, V. R. Paramathayalan, and P. Moulin, "Multiple granularity analysis for fine-grained action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 756–763, 2014.

[34] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289, 2001.

[35] T. Lan, T.-C. Chen, and S. Savarese, "A hierarchical representation for future action prediction," in *Proceedings of the European Conference on Computer Vision*, pp. 689–704, Springer, 2014.

[36] Y. Kong, D. Kit, and Y. Fu, "A discriminative model with multiple temporal scales for action prediction," in *Proceedings of the European Conference on Computer Vision*, pp. 596–611, Springer, 2014.

[37] N. Robertson and I. Reid, "A general method for human activity recognition in video," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 232–248, 2006.

[38] Y. Wang and G. Mori, "Learning a discriminative hidden part model for human action recognition," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1721–1728, 2009.

[39] Y. Song, L.-P. Morency, and R. Davis, "Action recognition by hierarchical sequence summarization," in *Proceedings of IEEE Conference in International Conference on Computer Vision*, (Portland, OR), pp. 3562–3569, Jun 2013.

[40] N. M. Oliver, B. Rosario, and A. P. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, 2000.

[41] Z. Wang, J. Wang, J. Xiao, K.-H. Lin, and T. Huang, "Substructure and boundary modeling for continuous action recognition," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 1330–1337, IEEE, 2012.

[42] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S. Chun Zhu, "Joint inference of groups, events and human roles in aerial videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4576–4584, 2015.

[43] C. Wu, J. Zhang, S. Savarese, and A. Saxena, "Watch-n-patch: Unsupervised understanding of actions and relations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4362–4370, 2015.

[44] W. Zhou and Z. Zhang, "Human action recognition with multiple-instance

markov model," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 10, pp. 1581–1591, 2014.

[45] W. Chen, C. Xiong, R. Xu, and J. J. Corso, "Actionness ranking with lattice conditional ordinal random fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 748–755, 2014.

[46] Y. Kong and Y. Fu, "Modeling supporting regions for close human interaction recognition.," in *Proceedings of the ECCV Workshops (2)*, pp. 29–44, 2014.

[47] A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2007.

[48] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *Proceedings of European Conference on Computer Vision*, (Crete, GRC), pp. 494–507, Springer, Sep 2010.

[49] Y. Wang and G. Mori, "A discriminative latent model of object classes and attributes," *Computer Vision–ECCV 2010*, pp. 155–168, 2010.

[50] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Proceedings of the Advances in neural information processing systems*, pp. 1410–1418, 2009.

[51] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 951–958, IEEE, 2009.

[52] V. I. Morariu and L. S. Davis, "Multi-agent event recognition in structured scenarios," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 3289–3296, IEEE, 2011.

[53] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 3337–3344, IEEE, 2011.

[54] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1331–1338, IEEE, 2011.

[55] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Attribute regularization based human action recognition," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 10, pp. 1600–1609, 2013.

[56] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rosenberg, and L. Fei-Fei, "Learning semantic relationships for better action retrieval in images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1100–1109, 2015.

[57] C.-Y. Chen and K. Grauman, "Efficient activity detection with max-subgraph search," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 1274–1281, IEEE, 2012.

[58] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 780–787, 2014.

[59] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proceedings of the International Conference on Pattern Recognition*, vol. 3, pp. 32–36, IEEE, 2004.

[60] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, pp. 1395–1402, IEEE, 2005.

[61] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of the Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pp. 65–72, IEEE, 2005.

[62] M. Jain, J. C. van Gemert, and C. G. Snoek, "What do 15,000 object categories tell us about classifying and localizing actions?," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 46–55, 2015.

[63] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[64] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1–8, Ieee, 2007.

[65] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.

[66] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proceedings of IEEE Conference in International Conference on Computer Vision*, (Miami, FL), pp. 1932–1939, IEEE, Jun 2009.

[67] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 444–451, IEEE, 2009.

[68] A. Oikonomopoulos, M. Pantic, and I. Patras, "Sparse b-spline polynomial descriptors for human activity recognition," *Image and Vision Computing*, vol. 27, no. 12, pp. 1814–1825, 2009.

[69] Y. Tian, R. Sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2642–2649, 2013.

[70] M. Jain, J. Van Gemert, H. Jégou, P. Bouthemy, and C. G. Snoek, "Action localization with tubelets from motion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 740–747, 2014.

[71] G. Yu, J. Yuan, and Z. Liu, "Propagative hough voting for human activity recognition," *Computer Vision–ECCV 2012*, pp. 693–706, 2012.

[72] M. Vrigkas, V. Karavasilis, C. Nikou, and I. A. Kakadiaris, "Matching mixtures of curves for human action recognition," *Computer Vision and Image Understanding*, vol. 119, pp. 27–40, 2014.

[73] B. Ni, P. Moulin, X. Yang, and S. Yan, "Motion part regularization: Improving action recognition via trajectory selection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3698–3706, 2015.

[74] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Proceedings of IEEE Conference in International Conference on Computer Vision*, (Portland, OR), pp. 2555–2562, Jun 2013.

[75] S. Samanta and B. Chanda, "Space-time facet model for human activity classification," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1525–1535, 2014.

[76] M. J. Roshtkhari and M. D. Levine, "Human activity recognition in videos using a single example," *Image and Vision Computing*, vol. 31, no. 11, pp. 864–876, 2013.

[77] Z. Jiang, Z. Lin, and L. S. Davis, "A unified tree-based framework for joint action localization, recognition and segmentation," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1345–1355, 2013.

[78] A. Gaidon, Z. Harchaoui, and C. Schmid, "Activity representation with motion hierarchies," *International Journal of Computer Vision*, vol. 107, no. 3, pp. 219–238, 2014.

[79] M. Raptis, I. Kokkinos, and S. Soatto, "Discovering discriminative action parts from mid-level video representations," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 1242–1249, IEEE, 2012.

[80] G. Yu and J. Yuan, "Fast action proposals for human action detection and search," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 1302–1311, 2015.

[81] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 1234–1241, IEEE, 2012.

[82] X. Yan and Y. Luo, "Recognizing human actions using a new descriptor based on spatial–temporal interest points and weighted-output classifier," *Neurocomputing*, vol. 87, pp. 51–61, 2012.

[83] B. Li, O. I. Camps, and M. Sznaier, "Cross-view activity recognition using hankelets," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 1362–1369, IEEE, 2012.

[84] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 489–496, IEEE, 2011.

[85] M. B. Holte, B. Chakraborty, J. Gonzalez, and T. B. Moeslund, "A local 3-d motion descriptor for multi-view human action recognition from 4-d spatio-temporal interest points," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 553–565, 2012.

[86] Q. Zhou and G. Wang, "Atomic action features: A new feature for action

recognition," in *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pp. 291–300, Springer, 2012.

[87] J. Sanchez-Riera, J. Čech, and R. Horaud, "Action recognition robust to background clutter by using stereo vision," in *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pp. 332–341, Springer, 2012.

[88] S. Khamis, V. I. Morariu, and L. S. Davis, "A flow model for joint action recognition and identity maintenance," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 1218–1225, IEEE, 2012.

[89] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 2855–2862, IEEE, 2012.

[90] M. Hoai, Z.-Z. Lan, and F. De la Torre, "Joint segmentation and classification of human actions in video," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 3265–3272, IEEE, 2011.

[91] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576–1588, 2012.

[92] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzàlez, "Selective spatio-temporal interest points," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 396–410, 2012.

[93] S. Satkin and M. Hebert, "Modeling the temporal extent of actions," *Computer Vision–ECCV 2010*, pp. 536–548, 2010.

[94] S. Ma, L. Sigal, and S. Sclaroff, "Space-time tree ensemble for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5024–5032, 2015.

[95] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proceedings of the*

IEEE Conference in Computer Vision and Pattern Recognition, pp. 2046–2053, IEEE, 2010.

[96] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 1290–1297, IEEE, 2012.

[97] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE Journal of selected topics in signal processing*, vol. 6, no. 5, pp. 538–552, 2012.

[98] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[99] M. Fergie and A. Galata, "Mixtures of gaussian process models for human pose estimation," *Image and Vision Computing*, vol. 31, no. 12, pp. 949–957, 2013.

[100] N. Ikizler and P. Duygulu, "Human action recognition using distribution of oriented rectangular patches," *Human Motion–Understanding, Modeling, Capture and Animation*, pp. 271–284, 2007.

[101] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1691–1703, 2012.

[102] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 412–424, 2012.

[103] M. Andriluka and L. Sigal, "Human context: Modeling human-human interac-

tions for monocular 3d pose estimation," *Articulated Motion and Deformable Objects*, pp. 260–272, 2012.

[104] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele, "Multi-view pictorial structures for 3d human pose estimation.," in *Proceedings of the British Machine Vision Conference*, 2013.

[105] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3d pictorial structures for multiple human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1669–1676, 2014.

[106] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, 2014.

[107] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1293–1301, 2015.

[108] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1110–1118, 2015.

[109] H. Yub Jung, S. Lee, Y. Seok Heo, and I. Dong Yun, "Random tree walk toward instantaneous 3d human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2467–2474, 2015.

[110] C. Thurau and V. Hlavác, "Pose primitive based human action recognition in videos or still images," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.

[111] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 3177–3184, IEEE, 2011.

[112] K. N. Tran, I. A. Kakadiaris, and S. K. Shah, "Part-based motion descriptor image for human action recognition," *Pattern Recognition*, vol. 45, no. 7, pp. 2562–2572, 2012.

[113] S. Sedai, M. Bennamoun, and D. Q. Huynh, "Discriminative fusion of shape and appearance features for human pose estimation," *Pattern Recognition*, vol. 46, no. 12, pp. 3223–3237, 2013.

[114] L. Sigal, M. Isard, H. Haussecker, and M. J. Black, "Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation," *International Journal of Computer Vision*, vol. 98, no. 1, pp. 15–48, 2012.

[115] A. Moutzouris, J. Martinez-del Rincon, J.-C. Nebel, and D. Makris, "Efficient tracking of human poses using a manifold hierarchy," *Computer Vision and Image Understanding*, vol. 132, pp. 75–86, 2015.

[116] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 588–595, 2014.

[117] I. Kviatkovsky, E. Rivlin, and I. Shimshoni, "Online action recognition using covariance of shape and motion," *Computer Vision and Image Understanding*, vol. 129, pp. 15–26, 2014.

[118] R. Anirudh, P. Turaga, J. Su, and A. Srivastava, "Elastic functional coding of human actions: From vector-fields to latent variables," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 3147–3155, 2015.

[119] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "Real time action recognition using histograms of depth gradients and random decision forests," in *Proceedings of the Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pp. 626–633, IEEE, 2014.

[120] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3361–3368, IEEE, 2011.

[121] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[122] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Proceedings of the International Workshop on Human Behavior Understanding*, pp. 29–39, Springer, 2011.

[123] H.-J. Kim, J. Lee, and H.-S. Yang, "Human action recognition using a modified convolutional neural network," *Advances in Neural Networks–ISNN 2007*, pp. 715–723, 2007.

[124] G. Gkioxari and J. Malik, "Finding action tubes," in *Proceedings of IEEE Conference in International Conference on Computer Vision*, (Boston, MA), pp. 759–768, Jun 2015.

[125] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4305–4314, 2015.

[126] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.

[127] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 166–173, IEEE, 2005.

[128] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *Pro-*

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 405–412, IEEE, 2005.

[129] T.-H. Yu, T.-K. Kim, and R. Cipolla, "Real-time action recognition by spatiotemporal semantic and structural forests.," in *Proceedings of the British Machine Vision Conference*, p. 6, 2010.

[130] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," *Computing Research Repository (CoRR)*, vol. abs/1604.06573, Apr 2016.

[131] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489–4497, 2015.

[132] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *Proceedings of IEEE Conference in International Conference on Computer Vision*, (Boston, MA), pp. 5378–5387, Jun 2015.

[133] Y. Wang and G. Mori, "Hidden part models for human action recognition: Probabilistic versus max margin," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1310–1323, 2011.

[134] D. Wu and L. Shao, "Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition," in *Proceedings of IEEE Conference in International Conference on Computer Vision*, (Columbus, OH), pp. 724–731, Jun 2014.

[135] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, "Script data for attribute-based recognition of composite activities," in *Proceedings of the European Conference on Computer Vision*, (Florence, ITA), pp. 144–157, Springer, Oct 2012.

[136] S. Bhattacharya, M. Kalayeh, R. Sukthankar, and M. Shah, "Recognition of complex events: Exploiting temporal dynamics between underlying concepts," in *Proceedings of IEEE Conference in International Conference on Computer Vision*, (Columbus, OH), pp. 2235–2242, Jun 2014.

[137] W. Li, Q. Yu, H. Sawhney, and N. Vasconcelos, "Recognizing activities via bag of words for attribute dynamics," in *Proceedings of IEEE Conference in International Conference on Computer Vision*, (Portland, OR), pp. 2587–2594, Jun 2013.

[138] A. Jackson, *Perspectives of Nonlinear Dynamics*, vol. 1. CUP Archive, 1992.

[139] T. Kailath, "A view of three decades of linear filtering theory," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 146–181, 1974.

[140] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of IEEE Conference in International Conference on Computer Vision*, (Boston, MA), pp. 4694–4702, Jun 2015.

[141] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of IEEE Conference in International Conference on Computer Vision*, (Boston, MA), pp. 2625–2634, Jun 2015.

[142] S. Ramasinghe and R. Rodrigo, "Action recognition by single stream convolutional neural networks: An approach using combined motion and static information," in *Proceedings of the Asian Conference on Pattern Recognition*, pp. 101–105, Nov 2015.

[143] L. Wang, Y. Qiao, and X. Tang, "Action recognition and detection by combining motion and appearance features," *THUMOS14 Action Recognition Challenge*, vol. 1, no. 2, p. 2, 2014.

[144] M. Xin, H. Zhang, M. Sun, and D. Yuan, "Recurrent temporal sparse autoencoder for attention-based action recognition," in *Neural Networks (IJCNN), 2016 International Joint Conference on*, pp. 456–463, IEEE, 2016.

[145] M. Xin, H. Zhang, H. Wang, M. Sun, and D. Yuan, "Arch: Adaptive recurrent-convolutional hybrid networks for long-term action recognition," *Neurocomputing*, vol. 178, pp. 87–102, 2016.

[146] G. Thung and H. Jiang, "A torch library for action recognition and detection using cnns and lstms,"

[147] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1942–1950, 2016.

[148] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," *Image analysis*, pp. 363–370, 2003.

[149] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2012.

[150] B. D. Lucas, T. Kanade, *et al.*, "An iterative image registration technique with an application to stereo vision," *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 1981.

[151] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1996–2003, IEEE, 2009.

[152] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," *Computer Vision-ECCV 2004*, pp. 25–36, 2004.

[153] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[154] D. Connaghan, C. . Conaire, P. Kelly, and N. E. O'Connor, "Recognition of tennis strokes using key postures," in *IET Irish Signals and Systems Conference (ISSC 2010)*, pp. 245–248, June 2010.

[155] E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot, "Hmm based structuring of tennis videos using visual and audio cues," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 3, pp. III–309, IEEE, 2003.

[156] C. Ó Conaire, D. Connaghan, P. Kelly, N. E. O'Connor, M. Gaffney, and J. Buckley, "Combining inertial and visual sensing for human action recognition in tennis," in *Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams*, pp. 51–56, ACM, 2010.

[157] T. Bloom and A. P. Bradley, "Player tracking and stroke recognition in tennis video," in *Proceedings of the APRS Workshop on Digital Image Computing (WDIC'03)*, vol. 1, pp. 93–97, The University of Queensland, 2003.

[158] G. Sudhir, J. C.-M. Lee, and A. K. Jain, "Automatic classification of tennis video for high-level content-based retrieval," in *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, pp. 81–90, IEEE, 1998.

[159] H. Miyamori and S.-I. Iisaku, "Video annotation for content-based retrieval using human behavior analysis and domain knowledge," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pp. 320–325, IEEE, 2000.

[160] Y. Gong, L. T. Sin, C. H. Chuan, H. Zhang, and M. Sakauchi, "Automatic pars-

ing of tv soccer programs," in *Multimedia Computing and Systems, 1995., Proceedings of the International Conference on*, pp. 167–174, IEEE, 1995.

[161] G. S. Pingali, Y. Jean, and I. Carlbom, "Real time tracking for enhanced tennis broadcasts," in *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp. 260–265, IEEE, 1998.

[162] G. Zhu, C. Xu, Q. Huang, W. Gao, and L. Xing, "Player action recognition in broadcast tennis video with applications to semantic analysis of sports game," in *Proceedings of the 14th ACM international conference on Multimedia*, pp. 431–440, ACM, 2006.

[163] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[164] S. Xiang, F. Nie, and C. Zhang, "Learning a mahalanobis distance metric for data clustering and classification," *Pattern Recognition*, vol. 41, no. 12, pp. 3600–3612, 2008.

[165] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei, "Imagenet large scale visual recognition competition ILSVRC," 2012.

[166] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proceedings of IEEE Conference in International Conference on Computer Vision*, (Miami, FL), pp. 2929–2936, IEEE, Jun 2009.

[167] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011.

[168] E. Vig, M. Dorr, and D. Cox, "Space-variant descriptor sampling for action recognition based on saliency and eye movements," in *Proceedings of European Conference on Computer Vision*, (Florence, ITA), pp. 84–97, Springer, Oct 2012.

[169] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo, "Trajectory-based modeling of human actions with motion reference points," in *Proceedings of European Conference on Computer Vision*, (Florence, ITA), pp. 425–438, Springer, Oct 2012.

[170] S. Mathe and C. Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in *Proceedings of European Conference on Computer Vision*, pp. 842–856, Florence, ITA: Springer, Oct 2012.

[171] Y. Zhu and S. D. Newsam, "Depth2action: Exploring embedded depth for large-scale action recognition," *CoRR*, vol. abs/1608.04339, 2016.

[172] M. M. Ullah, S. N. Parizi, and I. Laptev, "Improving bag-of-features action recognition with non-local cues.," in *Proceedings of British Machine Vision Conference*, vol. 10, pp. 95–1, 2010.

[173] H. Possegger, T. Mauthner, and H. Bischof, "In defense of color-based model-free tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2113–2120, 2015.

[174] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[175] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Proceedings of the International Conference on Database Theory*, pp. 420–434, Springer, 2001.

[176] P. Fränti, O. Virmajoki, and V. Hautamäki, "Fast agglomerative clustering using a k-nearest neighbor graph," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1875–1881, 2006.