ICASSP 2017 Turorial: Multimodal Signal Processing, Saliency and Summarization List of references

Petros Maragos, Alexandros Potamianos, Athanasia Zlatintsi and Petros Koutras

Sunday, March 5, 2017, 13:30 - 17:00

1 Multimodal Signal Processing, Audio-Visual Perception and Fusion

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

[2] Z. Barzelay and Y. Y. Schechner. Harmony in motion. 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. 25:828–836, 2003.

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

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

[7] T. Darrell, J. Fisher, P. Viola, and B. Freeman. Audio-visual segmentation and the cocktail party effect. 2000.

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

[9] Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Oscar Lopes, Isabelle Guyon, Vassilis Athitsos, and Hugo 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.

[10] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013.

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

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

[13] J. Hershey and J. Movellan. Audio-vision: Using audio-visual synchrony to locate sounds. 1999.

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

[15] Aggelos K Katsaggelos, Sara Bahaadini, and Rafael Molina. Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE*, 103(9):1635–1653, 2015.

[16] Athanassios Katsamanis, George Papandreou, and Petros 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.

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

[18] E. Kidron, Y. Y. Schechner, and M. Elad. Cross-modal localization via sparsity. 55(4):1390–1404, April 2007.

[19] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. 20(3):226–239, March 1998.

[20] 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.

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

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

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

[24] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

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

[26] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. A generic framework of user attention model and its application in video summarization. 7:907–919, 2005.

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

[28] Petros Maragos, Patrick Gros, Athanassios Katsamanis, and George Papandreou. Cross-modal integration for performance improving in multimedia: a review. In *Multimodal processing and interaction*, pages 1–46. Springer, 2008.

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

[30] Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Dong Zhang. Automatic analysis of multimodal group actions in meetings. 27:305–317, 2005.

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

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

[33] G. Monaci and P. Vandergheynst. Audiovisual Gestalts. page 200, New York, NY, 2006. IEEE Computer Society.

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

[35] George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis, and Petros 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.

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

[37] 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.

[38] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

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

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

[41] M.E. Sargin, Y. Yemez, E. Erzin, and A.M. Tekalp. Audiovisual synchronization and fusion using canonical correlation analysis. 9(7):1396–1403, November 2007.

[42] M. Slaney and M. Covell. FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks. 2001.

[43] Cees G.M. Snoek and Marcel Worring. Multimodal video indexing: A review of the state-of-the-art. 25(1):5–35, January 2005.

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

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

[46] Argiro Vatakis, Petros Maragos, Isidoros Rodomagoulakis, and Charles Spence. Assessing the effect of physical differences in the articulation of consonants and vowels on audiovisual temporal perception. *J Speech Lang Hear Res*, 2012.

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

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

[49] 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.

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

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

[52] 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.

2 Visual Processing and Saliency

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.

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

[3] Shumeet Baluja and Dean Pomerleau. Using a saliency map for active spatial selective attention: Implementation & initial results. In *Proc. NIPS*, 1994.

[4] Anna Belardinelli, Fiora Pirri, and Andrea Carbone. Motion saliency maps from spatiotemporal filtering. In *Attention in Cognitive Systems*, pages 112–123. Springer, 2009.

[5] Peng Bian and Liming Zhang. Biological plausibility of spectral domain approach for spatiotemporal visual saliency. In *Advances in Neuro-Information Processing*, volume 5506 of *Lecture Notes in Computer Science*, pages 251–258. 2009.

[6] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

[7] Ali Borji, Dicky N Sihite, and Laurent Itti. Probabilistic learning of task-specific visual attention. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 470–477, 2012.

[8] Ali Borji, Dicky N. Sihite, and Laurent Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Trans. Image Processing*, 22(1):55–69, 2013.

[9] A. C. Bovik, M. Clark, and W. S. Geisler. Multichannel texture analysis using localized spatial filters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(1):55–73, 1990.

[10] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *Proc. NIPS*, 2005.

[11] N. D. B. Bruce and J. K. Tsotsos. Spatiotemporal saliency: Towards a hierarchical representation of visual saliency. In *Int'l Workshop on Attention and Performance in Comp. Vis.*, 2008.

[12] Xinyi Cui, Qingshan Liu, and Dimitris Metaxas. Temporal spectral residual: Fast motion saliency detection. In *Proc. ACM Int'l Conf. on Multimedia*, 2009.

[13] John Daugman. Uncertainty Relation for Resolution in Space, Spatial Frequency and Orientation Optimized by Two-Dimensional Visual Cortical Filters. *J. Opt. Soc. Amer. A*, 2(7):1160–1169, 1985.

[14] John G Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20(10):847–856, 1980.

[15] Konstantinos G. Derpanis, Mikhail Sizintsev, Kevin Cannons, and Richard P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[16] 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, textual attention. *IEEE Trans. on Multimedia*, 15(7):1553–1568, Nov. 2013.

[17] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis. Video event detection and summarization using audio, visual and text saliency. In *Proc. Int'l Conf. Acoustics, Speech and Signal Processing*, 2009.

[18] Simone Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, volume 3899 of *Lecture Notes in Computer Science*. Springer, 2006.

[19] D. Gao, S. Han, and N. Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(6):989–1005, 2009.

[20] D. Gao and N. Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *Proc. NIPS*, 2004.

[21] Anton Garcia-Diaz, Xose R. Fernandez-Vidal, Xose Manuel Pardo, and Raquel Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51–64, 2012.

[22] Christos Georgakis, Petros Maragos, Georgios Evangelopoulos, and Dimitrios 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.

[23] Ioannis Gkioulekas, Georgios Evangelopoulos, and Petros Maragos. Spatial bayesian surprise for image saliency and quality assessment. In *Proc. Int'l Conf. Image Processing*, Sep. 2010.

[24] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[25] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans. Image Processing*, 19(1):185–198, 2010.

[26] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *NIPS*, 2006.

[27] Joseph P. Havlicek, David S. Harding, and Alan C. Bovik. Multidimensional quasi-eigenfunction approximations and multicomponent am-fm models. *IEEE Trans. Image Processing*, 9(2):227–242, 2000.

[28] David J. Heeger. Model for the extraction of image flow. *J. Opt. Soc. Amer.*, 4(8):1455–1471, 1987.

[29] David J. Heeger. Optical flow using spatio-temporal filters. *Int'l. J. Comput. Vis.*, 1(4):279–302, 1988.

[30] Xiaodi Hou, Jonathan Harel, and Christof Koch. Image signature: Highlighting sparse salient regions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(1):194–201, 2012.

[31] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[32] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: searching for coding length increments. In *NIPS*, 2009.

[33] L. Itti, N. Dhavale, and F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology*, 2003.

[34] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. In *Proc. NIPS*, 2005.

[35] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[36] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *Proc. Int'l Conf. on Computer Vision*, 2009.

[37] T. Kadir and M. Brady. Saliency, scale and image description. *Int'l. J. Comput. Vis.*, 45(2):83–105, 2001.

[38] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227, Jun 1985.

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

[40] Petros Koutras, Athanasia Zlatintsi, Elias Iosif, Athanasios Katsamanis, Petros Maragos, and Alexandros Potamianos. Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 4361–4365. IEEE, 2015.

[41] Se-Ho Lee, Jin-Hwan Kim, Kwang Pyo Choi, Jae-Young Sim, and Chang-Su Kim. Video saliency detection based on spatiotemporal feature learning. In *Proc. Int'l Conf. Image Processing*, 2014.

[42] Wen-Fu Lee, Tai-Hsiang Huang, Su-Ling Yeh, and Homer H Chen. Learning-based prediction of visual attention for video signals. *IEEE Trans. Image Processing*, 20(11):3028–3038, 2011.

[43] Jia Li, Yonghong Tian, Tiejun Huang, and Wen Gao. Probabilistic multi-task learning for visual saliency estimation in video. *Int. J. Comput. Vis.*, 90(2):150–165, 2010.

[44] Zhuang Li, Prakash Ishwar, and Janusz Konrad. Video condensation by ribbon carving. *IEEE Trans. on Image Processing*, 18(11):2572–2583, 2009.

[45] Zhi Liu, Wenbin Zou, Lina Li, Liquan Shen, and Olivier Le Meur. Co-saliency detection based on hierarchical segmentation. *IEEE Signal Processing Letters*, 21(1):88–92, 2014.

[46] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(1):171–177, 2010.

[47] M. Mancas, N. Riche, J. Leroy, and B. Gosselin. Abnormal motion selection in crowds using bottom-up saliency. In *Proc. Int'l Conf. Image Processing*, 2011.

[48] Kevis Maninis, Petros Koutras, and Petros 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.

[49] Sophie Marat, Tien Ho-Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, and Anne Guérin-Dugué. Modelling spatio-temporal saliency to predict gaze direction for short videos. *Int'l. J. Comput. Vis.*, 82(3):231–243, 2009.

[50] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483 – 2498, 2007.

[51] Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Pattern Analysis and Machine Intelligence*, 28(5):802–817, 2006.

[52] R. Milanese. *Detecting Salient Regions in an Image: From Biological Evidence to Computer Implementation*. PhD thesis, University of Geneva, 1993.

[53] Tam V Nguyen, Mengdi Xu, Guangyu Gao, Mohan Kankanhalli, Qi Tian, and Shuicheng Yan. Static saliency vs. dynamic saliency: a comparative study. In *Proc. ACM Int'l Conf. on Multimedia*, 2013.

[54] Ernst Niebur and Christof Koch. Control of selective visual attention: Modeling the where pathway. In *Proc. NIPS*, 1995.

[55] Aude Oliva, Antonio Torralba, Monica S. Castelhano, and John M. Henderson. Top-down control of visual attention in object detection. In *Proc. Int'l Conf. Image Processing*, 2003.

[56] Yael Pritch, Alex Rav-Acha, and Shmuel Peleg. Nonchronological video synopsis and indexing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(11):1971–1984, 2008.

[57] K. Rapantzikos, Y. Avrithis, and S. Kollias. Dense saliency-based spatiotemporal feature points for action recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.

[58] K. Rapantzikos, Y. Avrithis, and S. Kollias. Spatiotemporal features for action recognition and salient event detection. *Cognitive Computation, special issue on Saliency, attention, visual search and picture scanning*, 3(1):167–184, 2011.

[59] Nicolas Riche, Matei Mancas, Matthieu Duvinage, Makiese Mibulumukini, Bernard Gosselin, and Thierry Dutoit. Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, 28(6):642 – 658, 2013.

[60] Robert W Rodieck. Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vision research*, 5(12):583–601, 1965.

[61] Dmitry Rudoy, Dan B Goldman, Eli Shechtman, and Lihi Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[62] Boris Schauerte and Rainer Stiefelhagen. Quaternion-based spectral saliency detection for eye fixation prediction. In *Proc. Eur. Conf. Computer Vision*, 2012.

[63] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15, 2009.

[64] Alexander Toet. Computational versus Psychophysical Image Saliency: A Comparative Evaluation Study. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(11), 2011.

[65] Antonio Torralba. Modeling global scene factors in attention. *J. Opt. Soc. Amer. A*, 20:1407–1418, 2003.

[66] A.M. Treisman and G. Gelade. A feature integration theory of attention. *Cognit. Psychology*, 12(1):97–136, 1980.

[67] John K. Tsotsos, Sean M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, Oct. 1995.

[68] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *J. Neural Networks*, 19(9):1395–1407, 2006.

[69] R. A. Young, R. M. Lesperance, and W. W. Meyer. The Gaussian Derivative model for spatial-temporal vision: I. Cortical model. *Spatial Vision*, 14(3,4):261–319, 2001.

[70] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32.1–20, January 2008.

[71] Lingyun Zhang, Matthew H. Tong, and Garrison W. Sunday: Saliency using natural statistics for dynamic analysis of scenes. In *Proc. Thirty-first Annual Cognitive Science Society Conference.*, 2009.

[72] A. Zlatintsi, P. Maragos, A. Potamianos, and G. Evangelopoulos. A saliency-based approach to audio event detection and summarization. In *Proc. European Signal Process. Conf.*, 2012.

3 Audio Processing and Saliency

[1] Claude Alain and Lori J Bernstein. From sounds to meaning: the role of attention during auditory scene analysis. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 16(5):485–489, 2008.

[2] Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[3] Dimitrios Dimitriadis, Petros Maragos, and Alexandros Potamianos. On the effects of filterbank design and energy computation on robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1504–1516, 2011.

[4] Varinthira Duangudom and David V Anderson. Using auditory saliency to understand complex auditory scenes. In *Signal Processing Conference, 2007 15th European*, pages 1206–1210. IEEE, 2007.

[5] Georgios Evangelopoulos and Petros Maragos. Multiband modulation energy tracking for noisy speech detection. *IEEE Trans. Audio, Speech & Language Processing*, 14(6):2024–2038, 2006.

[6] Simon Haykin and Zhe Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005.

[7] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[8] James F Kaiser. On a simple algorithm to calculate the'energy'of a signal. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 381–384. IEEE, 1990.

[9] Ozlem Kalinli and Shrikanth S Narayanan. A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In *INTERSPEECH*, pages 1941–1944, 2007.

[10] Emine Merve Kaya and Mounya Elhilali. A temporal saliency map for modeling auditory attention. In *Information Sciences and Systems (CISS), 2012 46th Annual Conference on*, pages 1–6. IEEE, 2012.

[11] Emine Merve Kaya and Mounya Elhilali. Investigating bottom-up auditory attention. *Frontiers in human neuroscience*, 8:327, 2014.

[12] Christoph Kayser, Christopher I Petkov, Michael Lippert, and Nikos K Logothetis. Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology*, 15(21):1943–1947, 2005.

[13] Kyungtae Kim, Kai-Hsiang Lin, Dirk B Walther, Mark A Hasegawa-Johnson, and Tomas S Huang. Automatic detection of auditory salience with optimized linear filters derived from human annotation. *Pattern Recognition Letters*, 38:78–85, 2014.

[14] Petros Koutras, Athanasia Zlatintsi, Elias Iosif, Athanasios Katsamanis, Petros Maragos, and Alexandros Potamianos. Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 4361–4365. IEEE, 2015.

[15] Petros Maragos. Morphological filtering for image enhancement and feature detection. *analysis*, 19:18, 2005.

[16] Petros Maragos, James F Kaiser, and Thomas F Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE transactions on signal processing*, 41(10):3024–3051, 1993.

[17] Petros Maragos, Alex Potamianos, and Patrick Gros. *Multimodal processing and interaction: audio, video, text*, volume 33. Springer Science & Business Media, 2008.

[18] Reinier Plomp and Willem Johannes Maria Levelt. Tonal consonance and critical bandwidth. *The journal of the Acoustical Society of America*, 38(4):548–560, 1965.

[19] HM Teager and SM Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In *Speech production and speech modelling*, pages 241–261. Springer, 1990.

[20] Francesco Tordini, Albert S Bregman, Jeremy R Cooperstock, Anupryia Ankolekar, and Thomas Sandholm. Toward an improved model of auditory saliency. In *Proc. of the 19th Int. Conf. on Auditory Display (ICAD-2013)*. Georgia Institute of Technology, 2013.

[21] Tomoki Tsuchida and Garrison W Cottrell. Auditory saliency using natural statistics. In *CogSci*, 2012.

[22] Panteleimon Nestor Vassilakis. *Perceptual and physical properties of amplitude fluctuation and their musical significance*. PhD thesis, UNIVERSITY OF CALIFORNIA Los Angeles, 2001.

[23] Jingyu Wang, Ke Zhang, Kurosh Madani, and Christophe Sabourin. Salient environmental sound detection framework for machine awareness. *Neurocomputing*, 152:444–454, 2015.

[24] Athanasia Zlatintsi, Elias Iosif, Petros Marago, and Alexandros Potamianos. Audio salient event detection and summarization using audio and text modalities. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 2311–2315. IEEE, 2015.

[25] Athanasia Zlatintsi, Petros Maragos, Alexandros Potamianos, and Georgios Evangelopoulos. A saliency-based approach to audio event detection and summarization. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 1294–1298. IEEE, 2012.

[26] E Zwicker and H Fastl. *Psychoacoustics Facts and Models Springer Heiderberg*. Springer, 2nd edition, 1999.

4 Text Processing and Saliency

[1] Georgia Athanasopoulou, Elias Iosif, and Alexandros Potamianos. Low-dimensional manifold distributional semantic models. In *COLING*, pages 731–740, 2014.

[2] Frederic Charles Bartlett and Cyril Burt. Remembering: A study in experimental and social psychology. *British Journal of Educational Psychology*, 3(2):187–192, 1933.

[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[4] Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972.

[5] Stephen P. Borgatti. Identifying sets of key players in a network. In *Integration of Knowledge Intensive Multi-Agent Systems, 2003. International Conference on*, pages 127–131. IEEE, 2003.

[6] Jason M. Brenier, Daniel M. Cer, and Daniel Jurafsky. The detection of emphatic words using acoustic and lexical features. In *Interspeech*, pages 3297–3300, 2005.

[7] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[8] Moran Cerf, E. Paxon Frady, and Christof Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, 9(12):10–10, 2009.

[9] Allan M. Collins and Elizabeth F. Loftus. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407, 1975.

[10] John M. Conroy and Dianne P. O'leary. Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 406–407. ACM, 2001.

[11] Günes Erkan and Dragomir R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

[12] Linton C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[13] Sigmund Freud. Psychopathology of everyday life. 1938.

[14] Maria Fuentes, Enrique Alfonseca, and Horacio Rodríguez. Support vector machines for query-focused summarization trained and evaluated on pyramid data. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 57–60. Association for Computational Linguistics, 2007.

[15] Philip John Gorinski and Mirella Lapata. Movie script summarization as graph-based scene extraction. In *HLT-NAACL*, pages 1066–1076, 2015.

[16] Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204, 1986.

[17] Dilek Hakkani-Tur and Gokhan Tur. Statistical sentence extraction for information distillation. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1. IEEE, 2007.

[18] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[19] Julia Hirschberg. Pitch accent in context predicting intonational prominence from text. *Artificial Intelligence*, 63(1-2):305–340, 1993.

[20] Elias Iosif and Taniya Mishra. From speaker identification to affective analysis: A multi-step system for analyzing childrens stories. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature. Gothenburg, Sweden*, pages 40–49, 2014.

[21] Elias Iosif and Alexandros Potamianos. Feeling is understanding: From affective to semantic spaces. *IWCS 2015*, page 162, 2015.

[22] Valentin Jijkoun and Maarten De Rijke. Learning to transform linguistic graphs. In *Proc. of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing (TextGraphs-2)*, pages 53–60, 2007.

[23] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[24] Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73. ACM, 1995.

[25] Chin-Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 495–501. Association for Computational Linguistics, 2000.

[26] John Lyons. Ch. 15: Deixis, space and time, 1977.

[27] Nikolaos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2379–2392, 2013.

[28] Rada Mihalcea and Dragomir Radev. *Graph-based natural language processing and information retrieval*. Cambridge University Press, 2011.

[29] Rada Mihalcea and Paul Tarau. Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004.

[30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[31] Roberto Navigli and Mirella Lapata. Graph connectivity measures for unsupervised word sense disambiguation. In *IJCAI*, pages 1683–1688, 2007.

[32] Ani Nenkova, Kathleen McKeown, et al. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233, 2011.

[33] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[34] M. Ross Quillan. Semantic memory. Technical report, DTIC Document, 1966.

[35] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, pages 21–30. Association for Computational Linguistics, 2000.

[36] Kathleen Rastle, Matthew H. Davis, and Boris New. The broth in my brothers brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, 11(6):1090–1098, 2004.

[37] Henry L. Roediger and Kathleen B. McDermott. Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, 21(4):803, 1995.

[38] Timothy T. Rogers and James L. McClelland. *Semantic cognition: A parallel distributed processing approach*. MIT press, 2004.

[39] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[40] Martin Sarter, Ben Givens, and John P. Bruno. The cognitive neuroscience of sustained attention: where top-down meets bottom-up. *Brain research reviews*, 35(2):146–160, 2001.

[41] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[42] Peter D. Turney and Michael L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *arXiv preprint cs/0212012*, 2002.

[43] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

5 Multimodal Video Summarization

[1] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014.

[2] Sandra Eliza Fontes De Avila, Ana Paula Brandão Lopes, Antonio da Luz, and Arnaldo de Albuquerque Araújo. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68, 2011.

[3] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013.

[4] Georgios Evangelopoulos, Athanasia Zlatintsi, Georgios Skoumas, Konstantinos Rapantzikos, Alexandros Potamianos, Petros Maragos, and Y Avrithis. Video event detection and summarization using audio, visual and text saliency. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3553–3556. IEEE, 2009.

[5] Y. H. Gong and X. Liu. Video summarization using singular value decomposition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2000.

[6] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014.

[7] Michael Gygli, Helmut Grabner, and Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3090–3098, 2015.

[8] A. Hanjalic and H. J. Zhang. An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. *IEEE Trans. Circ. Syst. Video Technol.*, 9(8):1280–1289, 1999.

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

[10] Petros Koutras, Athanasia Zlatintsi, Elias Iosif, Athanasios Katsamanis, Petros Maragos, and Alexandros Potamianos. Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 4361–4365. IEEE, 2015.

[11] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011.

[12] Y. Ma, X.S. Hua, L. Lu, and H. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Trans. on Multimedia*, 7(5):907–919, Oct 2005.

[13] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. A generic framework of user attention model and its application in video summarization. *IEEE transactions on multimedia*, 7(5):907–919, 2005.

[14] Nikolaos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2379–2392, 2013.

[15] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2929–2936. IEEE, 2009.

[16] A Money and H Agius. Video summarization: A conceptual framework and survey of the state of the art. *J. Visual Communication and Image Representation*, 19(2):121–143, February 2008.

[17] Olivier Morère, Hanlin Goh, Antoine Veillard, Vijay Chandrasekhar, and Jie Lin. Co-regularized deep representations for video summarization. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 3165–3169. IEEE, 2015.

[18] Chong-Wah Ngo, Yu-Fei Ma, and Hong-Jiang Zhang. Automatic video summarization by graph modeling. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 104–109. IEEE, 2003.

[19] X. Orriols and X. Binefa. An EM algorithm for video summarization, generative model approach. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001.

[20] Georgia Panagiotaropoulou, Petros Koutras, Athanasios Katsamanis, Petros Maragos, Athanasia Zlatintsi, Athanassios Protopapas, Efstratios Karavasilis, and Nikolaos Smyrnis. Fmri-based perceptual validation of a computational model for visual and auditory saliency in videos. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 699–703. IEEE, 2016.

[21] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *European conference on computer vision*, pages 540–555. Springer, 2014.

[22] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044. ACM, 2014.

[23] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5179–5187, 2015.

[24] Y. Wang, Z. Liu, and J.-C. Huang. Multimedia content analysis using both audio and visual clues. *IEEE Signal Process. Mag.*, 17(6):12–36, November 2000.

[25] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1059–1067, 2016.

[26] Y. Zhuang, Y. Rui, T.S. Huang, and S. Mehrotra. Adaptive key frame extraction using unsupervised clustering. In *Proc. Int'l Conf. Image Processing*, 1998.

[27] Athanasia Zlatintsi, Elias Iosif, Petros Marago, and Alexandros Potamianos. Audio salient event detection and summarization using audio and text modalities. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 2311–2315. IEEE, 2015.

[28] Athanasia Zlatintsi, Petros Koutras, N Efthymiou, Petros Maragos, Alexandros Potamianos, and K Pastra. Quality evaluation of computational models for movie summarization. In *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, pages 1–6. IEEE, 2015.