个性化特征并综合分析不同学习者的共性特征,进而提供有针对性的评估、引导与干预.例如,在智慧课堂教学中,利用计算机实时分析统计学生行为,帮助老师及时掌握学生的学习特征和状态;在军训等集体活动中,预判学生可能发生的危险行为,提高安全防范能力;在中小学生课堂纪律维持方面,通过行为分析对学生的不良行为予以及时警告,避免其因课堂注意力不集中而导致学业警示等.

(2) 智能服务中的人机交互应用:有效的人机交互在任何服务型机器人应用中都至关重要.视觉场景描述技术提供了人机交互的自然语言交互接口.通过该技术,智能机器人能够以人类易于理解的自然语言方式来实现视觉场景内容信息的表达.另一方面,视频场景内容的自然语言描述也可以作为机器人内部场景的表现形式,为基于自然语言问答的智能环境感知提供了良好基础[76].使这些机器人可以像人一样有“感情”地进行语言表达,提供高质量的服务和陪伴是未来的研究重点之一.

(3) 视力障碍人员的辅助视听:该类应用旨在对人类活动场所中的视觉感知物体进行检测、识别、分析和判断,并给视力障碍人员予以提示,以辅助视力障碍人员顺利完成行为活动.其中,如何有效地将感知到的信息正确地传递给视力障碍人员是辅助视听应用技术的关键问题之一.如何快速、有效地感知人类活动场景中与活动相关的环境信息,通过视觉问答,并以友好的方式将相关信息传递给视力障碍人员是视觉场景描述应用中需解决的重要问题.

References:

[1] Wang J, Jiang W, Ma L, Liu W, Xu Y. Bidirectional attentive fusion with context gating for dense video captioning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 7190−7198.

[2] Ren Z, Wang X, Zhang N. Deep reinforcement learning-based image captioning with embedding reward. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1151−1159. [doi: 10.1109/CVPR.2017.128]

[3] Lu J, Xiong C, Parikh D, Socher R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 3242−3250. [doi: 10.1109/CVPR.2017.345]

[4] Wang Y, Lin Z, Shen X, Cohen S, Cottrell GW. Skeleton key: Image captioning by skeleton-attribute decomposition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 7272−7281. [doi: 10.1109/CVPR.2017.780]

[5] Pan Y, Yao T, Li H, Mei T. Video captioning with transferred semantic attributes. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 6504−6512. [doi: 10.1109/CVPR.2017.111]

[6] Zhang X, Gao K, Zhang Y, Zhang D, Li J, Tian Q. Task-driven dynamic fusion: Reducing ambiguity in video description. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 3713−3721. [doi: 10.1109/CVPR.2017.662]

[7] Shen Z, Li J, Su Z, Li M, Chen Y, Jiang YG, Xue XY. Weakly supervised dense video captioning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1916−1924. [doi: 10.1109/CVPR.2017.548]

[8] Krishna R, Hata K, Ren F, Niebles JC. Dense-captioning events in videos. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition 2017. 706−715. [doi: 10.1109/ICCV.2017.83]

[9] Huang Y, Wang W, Wang L. Instance-aware image and sentence matching with selective multimodal LSTM. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 7254−7262. [doi: 10.1109/CVPR.2017.767]

[10] Johnson J, Karpathy A, Li FF. DenseCap: Fully convolutional localization networks for dense captioning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4565−4574. [doi: 10.1109/CVPR.2016.494]

[11] Xu J, Mei T, Yao T, Rui Y. MSR-VTT: A large video description dataset for bridging video and language. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 5288−5296. [doi: 10.1109/CVPR.2016.571]

[12] Yu H, Wang J, Huang Z, Yang Y, Xu W. Video paragraph captioning using hierarchical recurrent neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4584−4593. [doi: 10.1109/CVPR.2016.496]

[13] You QZ, Jin HL, Wang ZW, Fang C, Luo JB. Image captioning with semantic attention. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4651−4659. [doi: 10.1109/CVPR.2016.503]

[14] Wu Q, Shen C, Liu L, Dick A, Hengel AVD. What value do explicit high level concepts have in vision to language problems. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 203−212. [doi: 10.1109/CVPR.2016.29]

[15] Devlin J, Cheng H, Fang H, Gupta S, Deng L, He XD, Zweig G, Mitchell Z. Language models for image captioning: The quirks and what works. In: Proc. of the Int'l Joint conf. on Natural Language Processing. 2015. 100–105. [doi: 10.3115/v1/P15-2017]

[16] Chen X, Zitnick CL. Mind's eye: A recurrent visual representation for image caption generation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 2422–2431. [doi: 10.1109/CVPR.2015.7298856]

[17] Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 3128–3137.

[18] Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 2625–2634. [doi: 10.1109/CVPR.2015.7298878]

[19] Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 4566–4575. [doi: 10.1109/CVPR.2015.7299087]

[20] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 3156–3164. [doi: 10.1109/CVPR.2015.7298935]

[21] Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, Courville AC. Describing videos by exploiting temporal structure. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 4507–4515. [doi: 10.1109/ICCV.2015.512]

[22] Das P, Xu C, Doell RF, Corso JJ. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2013. 2634–2641. [doi: 10.1109/CVPR.2013.340]

[23] Yao T, Pan Y, Li Y, Qiu Z, Mei T. Boosting image captioning with attributes. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 22–29. [doi: 10.1109/ICCV.2017.524]

[24] Chen TH, Liao YH, Chuang CY, Hsu WT, Fu JL, Sun M. Show, adapt and tell: Adversarial training of cross-domain image captioner. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 521–530. [doi: 10.1109/ICCV.2017.64]

[25] Li Y, Ouyang W, Zhou B. Scene graph generation from objects, phrases and region captions. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1261–1270. [doi: 10.1109/ICCV.2017.142]

[26] Na S, Lee S, Kim J, Kim G. A read-write memory network for movie story understanding. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 677–685. [doi: 10.1109/ICCV.2017.80]

[27] Hu R, Andreas J, Rohrbach M, Darrell T, Saenko K. Learning to reason: End-to-end module networks for visual question answering. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 804–813. [doi: 10.1109/ICCV.2017.93]

[28] Teney D, Liu L, Den Hengel AV. Graph-structured representations for visual question answering. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1536–1544. [doi: 10.1109/ICCV.2017.93]

[29] Zhu C, Zhao Y, Huang S. Structured attentions for visual question answering. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1291–1300. [doi: 10.1109/ICCV.2017.145]

[30] Xu H, Venugopalan S, Ramanishka V, Rohrbach M, Saenko K. A multi-scale multiple instance video description network. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 272–279.

[31] Venugopalan S, Rohrbach M, Donahue J. Mooney RJ, Darrell T, Saenko K. Sequence to sequence-video to text. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 4534–4542. [doi: 10.1109/ICCV.2015.515]

[32] Guadarrama S, Krishnamoorthy N, Malkarnenkar G, Venugopalan S, Mooney R, Darrell T, Saenko K. YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 2712–2719. [doi: 10.1109/ICCV.2013.337]

[33] Rohrbach M, Qiu W, Titov I. Translating video content to natural language descriptions. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 433–440. [doi: 10.1109/ICCV.2013.61]

[34] Mallya A, Lazebnik S. Learning models for actions and person-object interactions with transfer to question answering. In: Proc. of the European Conf. on Computer Vision. 2016. 414–428. [doi: 10.1007/978-3-319-46448-0_25]

[35] Rohrbach A, Rohrbach M, Hu R. Grounding of textual phrases in images by reconstruction. In: Proc. of the European Conf. on Computer Vision. 2016. 817–834. [doi: 10.1007/978-3-319-46448-0_49]

[36] Lu C, Krishna R, Bernstein M. Visual relationship detection with language priors. In: Proc. of the European Conf. on Computer Vision. 2016. 852−869. [doi: 10.1007/978-3-319-46448-0_51]

[37] Peter A, Basura F, Mark J, Stephen G. SPICE: Semantic propositional image caption evaluation. In: Proc. of the European Conf. on Computer Vision. 2016. 382−398.

[38] Lin X, Parikh D. Leveraging visual question answering for image-caption ranking. In: Proc. of the European Conf. on Computer Vision. 2016. 261−277. [doi: 10.1007/978-3-319-46475-6_17]

[39] Wu Q, Cai HP, Hall P. Learning graphs to model visual objects across different depictive styles. In: Proc. of the European Conf. on Computer Vision. 2014. 313−328. [doi: 10.1007/978-3-319-10584-0_21]

[40] Farhadi A, Hejrati M, Sadeghi MA. Every picture tells a story: Generating sentences from images. In: Proc. of the European Conf. on Computer Vision. 2010. 15−29. [doi: 10.1007/978-3-642-15561-1_2]

[41] Seo PH, Lehrmann A, Han B. Visual reference resolution using attention memory for visual dialog. In: Proc. of the 31st Annual Conf. on Neural Information Processing Systems. 2017. 3722−3732.

[42] Dai B, Lin D. Contrastive learning for image captioning. In: Proc. of the 31st Annual Conf. on Neural Information Processing Systems. 2017. 898−907.

[43] Wang L, Schwing A, Lazebnik S. Diverse and accurate image description using a variational auto-encoder with an additive Gaussian encoding space. In: Proc. of the 31st Annual Conf. on Neural Information Processing Systems. 2017. 5758−5768.

[44] Yeh R, Xiong J, Hwu WM. Interpretable and globally optimal prediction for textual grounding using image concepts. In: Proc. of the 31st Annual Conf. on Neural Information Processing Systems. 2017. 1909−1919.

[45] Fidler S. Teaching machines to describe images with natural language feedback. In: Proc. of the Neural Information Processing Systems. 2017. 5075−5085.

[46] Yang Z, Yuan Y, Wu Y. Review networks for caption generation. In: Proc. of the 30th Annual Conf. on Neural Information Processing Systems. 2016. 2361−2369.

[47] Fang H, Gupta S, Iandola F. From captions to visual concepts and back. In: Proc. of the 29th Annual Conf. on Neural Information Processing Systems. 2015. 1473−1482. [doi: 10.1109/CVPR.2015.7298754]

[48] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the 26th Annual Conf. on Neural Information Processing Systems. 2012. 1097−1105. [doi: 10.1145/3065386]

[49] Ordonez V, Kulkarni G, Berg TL. Im2text: Describing images using 1 million captioned photographs. In: Proc. of the 25th Annual Conf. on Neural Information Processing Systems. 2011. 1143−1151.

[50] Cho K, Van MB, Gulcehre C. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proc. of the 3rd Int'l Symp. on Natural Language Processing. 2014. 1724−1734. [doi: 10.3115/v1/D14-1179]

[51] Yang Y, Teo CL, Aloimonos Y. Corpus-guided sentence generation of natural images. In: Proc. of the Int'l Symp. Natural Language Processing. 2011. 444−454.

[52] Subhashini V, Xu HJ, Donahue J, Rohrbach M, Mooney R, Saenko K. Translating videos to natural language using deep recurrent neural networks. In: Proc. of the 2015 Conf. of the North American Chapter of the Association for Computational Linguistics. 2015. 1494−1504. [doi: 10.3115/v1/N15-1173]

[53] Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In: Proc. of the Int'l Conf. on Machine Learning. 2011. 689−696.

[54] Wang K, He R, Wang L, Wang W, Tan T. Joint feature selection and subspace learning for cross-modal retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2016,38(10):2010−2023. [doi: 10.1109/TPAMI.2015.2505311]

[55] Li XL, Shi JH, Dong YS, Tao DC. A survey on scene image classication. Scientia Sinica Informationis, 2015,45(7):827−848 (in Chinese with English abstract).

[56] Lowry SM, Sunderhauf N, Newman P, Leonard JJ, Cox DD, Corke P, Milford M. Visual place recognition: A survey. IEEE Trans. on Robotics, 2016,32(1):1−19. [doi: 10.1109/TRO.2015.2496823]

[57] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556[cs.CV], 2014.

[58] Song X, Jiang S, Herranz L. Multi-scale multi-feature context modeling for scene recognition in the semantic manifold. IEEE Trans. on Image Processing, 2017,26(6):2721−2735. [doi: 10.1109/TIP.2017.2686017]

[59] Luis H, Jiang SQ, Li XY. Scene recognition with CNNs: Objects, scales and dataset bias. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 571−579. [doi: 10.1109/CVPR.2016.68]

[60] Zhang H, Kyaw Z, Chang S, Chua T. Visual translation embedding network for visual relation detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 3107−3115. [doi: 10.1109/CVPR.2017.331]

[61] Wu Q, Shen C, Wang P, Dick A, Hengel AVD. Image captioning and visual question answering based on attributes and external knowledge. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017,40(6):1367−1381. [doi: 10.1109/TPAMI.2017.2708709]

[62] Johnson J, Krishna R, Stark M, Li L, Shamma DA, Bernstein MS, Feifei L. Image retrieval using scene graphs. In: Proc. of the Computer Vision and Pattern Recognition. 2015. 3668−3678. [doi: 10.1109/CVPR.2015.7298990]

[63] Li XY, Jiang SQ. Bundled object context for referring expressions. IEEE Trans. on Multimedia, 2018,20(10):2749−2760.

[64] Krahmer E, Van Deemter K. Computational generation of referring expressions: A survey. Computational Linguistics, 2012,38(1):173−218. [doi: 10.1162/COLI_a_00088]

[65] Mitchell M, Han X, Dodge J, Mensch A, Goyal A, Berg A, Yamaguchi K, Berg T, Stratos K, Daume H III. Midge: Generating image descriptions from computer vision detections. In: Proc. of the European Association of Computational Linguistics. 2012. 747−756.

[66] Krishnamoorthy N, Malkarnenkar G, Mooney R, Saenko K, Guadarrama S. Generating natural-language video descriptions using text-mined knowledge. In: Proc. of the Association for the Advance of Artificial Intelligence. 2013. 541−547.

[67] Kulkarni G, Premraj V, Dhar S, Berg AC, Berg TL. Babytalk: Understanding and generating simple image descriptions. IEEE Trans.on Pattern Analysis and Machine Intelligence, 2013,35(12):2891−2903. [doi: 10.1109/TPAMI.2012.162]

[68] Lebret R, Pinheiro PHO, Collobert R. Phrase-based image captioning. arXiv: 1502.03671[cs.CV], 2015.

[69] Kojima A, Tamura T, Fukunaga K. Natural language description of human activities from video images based on concept hierarchy of actions. Int'l Journal of Computer Vision, 2002,50(2):171−184.

[70] Xu R, Xiong C, Chen W, Corso JJ. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: Proc. of the Association for the Advance of Artificial Intelligence. 2015. 2346−2352.

[71] Kuznetsova P, Ordonez V, Berg T, Choi Y. Treetalk: Composition and compression of trees for image descriptions. Trans. of the Association of Computational Linguistics, 2014,2(10):351−362.

[72] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. In: Proc. of the Association for the Advance of Artificial Intelligence. 2015. 4188−4192. [doi: 10.1613/jair.3994]

[73] Devlin J, Gupta S, Girshick R, Mitchell M, Zitnick CL. Exploring nearest neighbor approaches for image captioning. arXiv: 1505.04467[cs.CV], 2015.

[74] Graves A. Supervised Sequence Labeling with Recurrent Neural Networks. Berlin, Heidelberg: Springer-Verlag, 2012.

[75] Lipton Z, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. arXiv: 1506.00019[cs.CV], 2015.

[76] Cascianelli S, Costante G, Ciarfuglia TA, Valigi P, Fravolini ML. Full-GRU natural language video description for service robotics applications. IEEE Robotics & Automation Letters, 2018,3(2):841−848. [doi: 10.1109/LRA.2018.2793345]

[77] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: Proc. of the Int'l Conf. on Machine Learning. 2015. 2048−2057.

[78] Socher R, Karpathy A, Le QV, Manning CD, Ng AY. Grounded compositional semantics for finding and describing images with sentences. Trans. of the Association for Computational Linguistics, 2014,2:207−218.

[79] Torabi A, Tandon N, Sigal L. Learning language-visual embedding for movie understanding with natural-language. arXiv: 1609.08124[cs.CV],2016.

[80] Everingham M, Zisserman A, Williams CKI, *et al*. The 2005 PASCAL visual object classes challenge. In: Proc. of the Int'l Conf. on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment. 2005. 117−176. [doi: 10.1007/11736790_8]

[81] Young P, Lai A, Hodosh M, Micah H, Julia H. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. of the Association for Computational Linguistics, 2014,2(1):67−78.

[82] Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. Int'l Journal of Computer Vision, 2015,123(1):74−93. [doi: 10.1007/s11263-016-0965-7]

[83] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollar P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the European Conf. on Computer Vision. 2014. 740−755. [doi: 10.1007/978-3-319-10602-1_48]

[84] Ni K, Pearce R, Boakye K, Van Essen B, Borth D, Chen B, Wang EX. Large-scale deep learning on the YFCC100M dataset. arXiv: 1502.03409[cs.CV], 2015.

[85] Krishna R, Zhu Y, Groth O, *et al*. Visual Genome: Connecting language and vision using crowd sourced dense image annotations. Int'l Journal of Computer Vision, 2017,123(1):32−73. [doi: 10.1007/s11263-016-0981-7]

[86] Wu J, Zheng H, Zhao B, Li YX, Yan BM, Liang R, *et al*. AI challenger: A large-scale dataset for going deeper in image understanding. arXiv: 1711.06475v1 [cs.CV], 2017.

[87] Chen DL, Dolan WB. Collecting highly parallel data for paraphrase evaluation. In: Proc. of the Association for Computational Linguistics: Human Language Technologies. 2011. 190−200.

[88] Rohrbach A, Rohrbach M, Qiu W, Friedrich A, Pinkal M, Schiele B. Coherent multi-sentence video description with variable level of detail. In: Proc. of the German Conf. on Pattern Recognition. 2014. 184−195. [doi: 10.1007/978-3-319-11752-2_15]

[89] Rohrbach A, Rohrbach M, Tandon N, Schiele B. A dataset for movie description. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 3202−3212. [doi: 10.1109/CVPR.2015.7298940]

[90] Torabi A, Pal C, Larochelle H, Courville A. Using descriptive video services to create a large data source for video annotation research. arXiv: 1503.01070v1[cs.CV], 2015.

[91] Zhou L, Xu C, Corso JJ. Towards automatic learning of procedures from Web instructional videos. In: Proc. of the Association for the Advance of Artificial Intelligence. 2018. 7591−7598.

[92] Papineni K, Roukos S, Ward T, Zhu W. BLEU: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Meeting on Association for Computational Linguistics. 2002. 311−318. [doi: 10.3115/1073083.1073135]

[93] Callison-Burch C, Osborne M, Koehn P. Re-evaluation the role of Bleu in machine translation research. In: Proc. of the European Association of Computational Linguistics. 2006. 249−256.

[94] Mahathir F. Sistem pendeteksi plagiat pada dokumen teks berbahasa Indonesia menggunakan metode Rouge-N, Rouge-L dan Rouge-W. 2011. http://repository.ipb.ac.id/handle/123456789/50046

[95] Lin CY, Och FJ. Automatic evaluation of machine translation quality using longest common subsequence and Skip-bigram statistics. In: Proc. of the Association for Computational Linguistics. 2004. 605−612. [doi: 10.3115/1218955.1219032]

[96] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proc. of the 43rd Annual Meeting on Association for Computational Linguistics. 2005,(29):65−72.

[97] Wong B, Kit C. ATEC: Automatic evaluation of machine translation via word choice and word order. Machine Translation, 2009, 23(2-3):141−155. [doi: 10.1007/s10590-009-9061-x]

[98] Kusner MJ, Sun Y, Kolkin NI, Weinberger KQ. From word embeddings to document distances. In: Proc. of the Int'l Conf. on Machine Learning. 2015. 957−966.

[99] Anderson P, Fernando B, Johnson M. SPICE: Semantic propositional image caption evaluation. In: Proc. of the European Conf. on Computer Vision. 2016. 382−398. [doi: 10.1007/978-3-319-46454-1_24]

[100] Ma M, Wang B. A grey relational analysis based evaluation metric for image captioning and video captioning. In: Proc. of the Grey Systems and Intelligent Services. 2017. 76−81. [doi: 10.1109/GSIS.2017.8077673]

[101] Cui Y, Yang G, Veit A, Huang X, Belongie SJ. Learning to evaluate image captioning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 5804−5812.

[102] Kilickaya M, Erdem A, Ikizler-Cinbis N, Erdem E. Re-evaluating automatic metrics for image captioning. In: Proc. of the European Chapter of the Association for Computational Linguistics. 2016. 199−209. [doi: 10.18653/v1/E17-1019]

[103] Giménez J, Màrquez L. Linguistic features for automatic evaluation of heterogenous MT systems. In: Proc. of the 2nd Workshop on Statistical Machine Translation. 2007. 256−264.

[104] ShafieiBavani E, Ebrahimi M, Wong R, Chen F. A semantically motivated approach to compute ROUGE scores. arXiv: 1710.07441 [cs.CV], 2017.

[105] Koehn P, Monz C. Manual and automatic evaluation of machine translation between European languages. In: Proc. of the Workshop on Statistical Machine Translation. 2006. 102−121. [doi: 10.3115/1654650.1654666]

附中文参考文献:

[55] 李学龙,史建华,董永生,陶大程.场景图像分类技术综述,中国科学:信息科学,2015,45(7):827−848.

马苗(1977－),女,山东聊城人,博士,教授,CCF 高级会员,主要研究领域为图像处理,模式识别,视频分析.

王伯龙(1993－),男,硕士,主要研究领域为视频分析与描述.

吴琦(1987－),男,博士,助理教授,博士生导师,主要研究领域为计算机视觉,机器学习,视觉问答.

武杰(1985－),男,博士,讲师,主要研究领域为遥感影像处理.

郭敏(1964－),女,博士,教授,博士生导师,主要研究领域为图像处理,模式识别,智能信息处理.