## Acknowledgments

This work was supported by DoD SBIR Award N00014-12-C-0263, the Google Faculty Research Award, NSF Award 1012017 and 1319378, ONR Award N00014-11-10417, ARO Award W911NF-08-1-0242, and the Intel Science & Technology Center for Pervasive Computing (ISTC-PC).

## References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Susstrunk, S. 2012. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(11):2274–2282.

Begley, C. G., and Ellis, L. M. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483(7391):531–533.

Branavan, S. R. K.; Chen, H.; Zettlemoyer, L. S.; and Barzilay, R. 2009. Reinforcement learning for mapping instructions to actions. In *ACL/AFNLP*, 82–90.

Brown, P. F.; Della Pietra, S. A.; Della Pietra, V. J.; and Mercer, R. L. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2):263–311.

Brown, P. F.; Lai, J. C.; and Mercer, R. L. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th ACL*, 169–176. ACL.

Charniak, E., and Johnson, M. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *ACL*.

Cour, T.; Jordan, C.; Miltsakaki, E.; and Taskar, B. 2008. Movie/script: Alignment and parsing of video and text transcription. In *Proceedings of the 10th European Conference on Computer Vision: Part IV*, ECCV '08, 158–171. Berlin, Heidelberg: Springer-Verlag.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society* 39(1):1–21.

Duygulu, P.; Batan, M.; and Forsyth, D. 2006. Translating images to words for recognizing objects in large image and video collections. In Ponce, J.; Hebert, M.; Schmid, C.; and Zisserman, A., eds., *Toward Category-Level Object Recognition*, volume 4170 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 258–276.

Duygulu, P.; Barnard, K.; Freitas, J. d.; and Forsyth, D. A. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision (ECCV)*, 97–112. Springer-Verlag.

Jamieson, M.; Dickinson, S.; Stevenson, S.; and Wachsmuth, S. 2006. Using language to drive the perceptual grouping of local image features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, volume 2, 2102–2109.

Kipp, M. 2012. Anvil: A universal video research tool. *Handbook of Corpus Phonology. Oxford University Press*.

Krishnamoorthy, N.; Malkarnenkar, G.; Mooney, R.; Saenko, K.; and Guadarrama, S. 2013. Generating natural-language video descriptions using text-mined knowledge. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-13)*, volume 2013, 3.

Krishnamurthy, J., and Kollar, T. 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Assoc. for Comp. Ling.* 10:193–206.

Lei, J.; Ren, X.; and Fox, D. 2012. Fine-grained kitchen activity recognition using rgb-d. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (Ubicomp)*, 208–211. ACM.

Li, Y., and Luo, J. 2013. Task-relevant object detection and tracking. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*.

Liang, P.; Jordan, M. I.; and Klein, D. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, 91–99. Association for Computational Linguistics.

Liang, P.; Jordan, M. I.; and Klein, D. 2011. Learning dependency-based compositional semantics. In *ACL*, 590–599.

Luo, J., and Guo, C. 2003. Perceptual grouping of segmented regions in color images. *Pattern Recognition* 36(12):2781 – 2792.

Matuszek, C.; Fitzgerald, N.; Zettlemoyer, L.; Bo, L.; and Fox, D. 2012. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML-2012)*, 1671–1678.

Moore, R. C. 2002. Fast and accurate sentence alignment of bilingual corpora. In *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, 135–144. London, UK: Springer-Verlag.

Rabiner, L. R. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

Rohrbach, M.; Qiu, W.; Titov, I.; Thater, S.; Pinkal, M.; and Schiele, B. 2013. Translating video content to natural language descriptions. In *14th IEEE International Conference on Computer Vision (ICCV)*, 433–440.

Song, S., and Xiao, J. 2013. Tracking revisited using rgbd camera: Unified benchmark and baselines. In *14th IEEE International Conference on Computer Vision (ICCV 2013)*. IEEE.

Tellex, S. A.; Kollar, T. F.; Dickerson, S. R.; Walter, M. R.; Banerjee, A.; Teller, S.; and Roy, N. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-11)*. AAAI Publications.

Tellex, S.; Thaker, P.; Joseph, J.; and Roy, N. 2013. Learning perceptually grounded word meanings from unaligned parallel data. *Machine Learning* 1–17.

Vogel, A., and Jurafsky, D. 2010. Learning to follow navigational directions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 806–814.

Vogel, S.; Ney, H.; and Tillmann, C. 1996. HMM-based word alignment in statistical translation. In *COLING-96*, 836–841.

Wachsmuth, S.; Stevenson, S.; and Dickinson, S. 2003. Towards a framework for learning structured shape models from text-annotated images. In *Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-linguistic Data - Volume 6*, HLT-NAACL-LWM '04, 22–29. Stroudsburg, PA, USA: Association for Computational Linguistics.

Yu, H., and Siskind, J. M. 2013. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-13)*, volume 1, 53–63.

Zettlemoyer, L. S., and Collins, M. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *UAI*, 658–666.

Zettlemoyer, L. S., and Collins, M. 2009. Learning context-dependent mappings from sentences to logical form. In *ACL/AFNLP*, 976–984.