The Semantics of Context: The Role of Interpretation and Belief in Visual Localization for Robots

Stephanie Lowry
Centre for Applied Autonomous Sensor Systems
Örebro University
70281 Örebro, Sweden
stephanie.lowry@oru.se

Abstract

While visual localization has improved in performance dramatically in recent years due to the development of ever-improving robust representations of locations, this paper considers a different aspect of the problem – the belief generation process. Belief generation is the conversion of a measure of similarity between two location representations to a measure of sameness – that is, are these two representations captured at the same location – which can be affected by the level of perceptual aliasing within the environment as well as the level of perceptual change. While probabilistic formulations of visual localization address these issues, environmental context can critically affect the performance of these belief generation methods.

1 Introduction

Visual localization is the process of using information about the external environment captured from vision-based sensors to correct a mobile autonomous system's belief about its location [LSN+16]. Visual localization systems have improved to an astonishing degree and modern systems are capable of impressive place recognition performance in extremely challenging conditions [SSD+15, AGT+18, KEC+18]. Much of this improvement is inspired by the rapid advances in machine learning for computer vision [LBH15] which allows localization systems to exploit sophisticated deep learning image representations. These robust representations can be used to compare two locations and determine if the appearance of these locations are similar.

A key challenge in visual localization is translating from the *similarity* of two locations to whether these two locations are actually the *same* place. Generally speaking, there are two scenarios that cause errors in visual localization: *perceptual aliasing* and *perceptual change*. These scenarios cause false positive and false negative errors respectively. Perceptual aliasing causes the system to believe observations from two separate locations are actually of the same place (false positive). Perceptual change causes observations of the same location not to be correctly identified as such (false negative).

Perceptual aliasing occurs when places in the environment look very similar and it is difficult to identify one location from another. Motorway driving can produce perceptual aliasing: motorways often look very similar over long distances, so you need to investigate highly distinctive elements – such as exit numbers on signs – to

know exactly where you are. Even if long stretches of the road appear superficially similar, you need to reject the possibility that they are actually the same place.

The second scenario is *perceptual change*: places in the environment do not look the same as they did on previous occasions. Perceptual change is particularly challenging when it happens uniformly over a region – for example, if day turns to night or snow falls – as all places in that region will become difficult to recognize. However, the spatial relationship between places does not change, so as long some similarity in appearance remains, a weak location hypothesis can be formed and by observing multiple nearby locations and the spatial relationship between them, a system can gradually build up confidence in its location belief.

There is an inherent conflict between resolving perceptual aliasing and perceptual change: perceptual aliasing requires adhering to a strict matching strategy where places must be both highly similar and highly distinctive, while perceptual change requires a permissive matching strategy where places may be matched together even when they do not appear similar at all. Thus an important consideration for a localization system is *context* – is the system in a situation when it should demand highly rigorous matching expectations or is a more permissive strategy necessary?

2 Belief Generation

As discussed in [LSN+16], there are a number of methods for determining whether two location representations were captured at the same location, such as voting methods [SBS07] and – when techniques inspired by text-based document analysis were used – the term frequency–inverse document frequency (TF-IDF) was used [NG12] to measure the mutual information between representations. A probabilistic formulation was also often used. Using a probabilistic framework has advantages: it provides a mechanism for managing uncertainty introduced from various sources and it naturally outputs a measure that expresses the degree of confidence in the current location belief. One well-known probabilistic framework for visual localization is FAB-MAP [CN08]. FAB-MAP uses a Naïve Bayes or a Chow-Liu [CL68] approximation to simplify the complex joint probability between the visual words in its model, and introduces a hidden variable to reduce the probabilities to quantities that can be calculated from training data.

An important aspect of location matching is geometric consistency – elements within the environment should stay in the same physical position relative to each other. Geometric verification tests can eliminate false positive matches using spectral clustering [Ols09] or RANSAC [FB81]. Furthermore, not only will elements within a location remain geometrically consistent, but the spatial relationship of the locations themselves will remain constant. These spatial relationships can be integrated into the localization belief probabilistically [BHK12] or via other methods such as network flows [NSBS14]. The spatial relationship between locations is extremely important information – in an extreme case, impressive localization results can still be achieved when appearance information is ignored and only odometry information is used [BGU16].

A number of trade-offs between competing priorities have been observed. A localization system becomes increasingly dependent on spatial information when there is extreme perceptual change [MW12, NSBS14, HB14], thus increasing sensitivity to motion uncertainty and reducing viewpoint flexibility. It has also been shown that methods that perform more effectively in perceptually changing environments do not provide as accurate localization [SMT+18]. These trade-offs suggest a choice must be made as to the requirements of the application and the operating environment.

3 Probabilistic Belief Generation

This section presents the formalism behind a probabilistic formulation of visual localization. Using a probabilistic framework has many advantages: it provides a mechanism for managing uncertainty introduced from various sources and naturally outputs a measure that expresses the degree of confidence in the current location belief. Probabilistic localization also naturally integrates some spatial environmental context; if there is a great deal of perceptual aliasing in the environment, the system's confidence in its location belief will be low.

The probabilistic framework is applied recursively over time as the system moves through the world: the prior belief is updated based on the system's motion model. The uncertainty in the system's location belief is continuously increased by the error in the motion model, and would grow in an unbounded manner if it were not constrained and corrected by the external observations of the world.

Formally, visual localization is probabilistically defined as follows (using the same notation as [CN08]): at time step k, the system has made a series of observations $\mathcal{Z}^k = \{Z_0, Z_1, \ldots, Z_k\}$, and has previously visited locations

$\{L_0, L_1, \ldots, L_{k-1}\}$. The likelihood of the system being in location L_i at time k given the current observation Z_k is

$$p(L_i \mid \mathcal{Z}^k) = \frac{p(Z_k \mid L_i, \mathcal{Z}^{k-1}) p(L_i \mid \mathcal{Z}^{k-1})}{p(Z_k \mid \mathcal{Z}^{k-1})}. \quad (1)$$

The second term in the numerator $p(L_i \mid \mathcal{Z}^{k-1})$ is the ***location prior***: the system's prior belief about the location before making the current observation Z_k. It allows the localization to be updated recursively over multiple timesteps: at time step $k+1$ the output of Equation 1 becomes the new location prior and Equation 1 can be applied again.

The other two terms are the ***observation likelihood models***. The first term in the numerator $p(Z_k \mid L_i, \mathcal{Z}^{k-1})$ is the likelihood that the robot would make observation Z_k if it is indeed at location L_i, and the denominator $p(Z_k \mid \mathcal{Z}^{k-1})$ is a normalizing factor that determines the likelihood that the robot would make the observation Z_k anywhere within the environment, thereby introducing further spatial context into the localization calculation.

4 Observation Likelihood Models

The observation likelihood models are the mechanisms by which context is introduced into the localization calculation. The system's observations naturally have a degree of uncertainty and have to be interpreted within the context of the environment: for example, should the system have a strict or a permissive matching strategy? Furthermore, the probabilistic framework does not implicitly ensure temporal environmental context (that is, perceptual change) is included.

The observation likelihood models themselves need to be learned or at least embody some data-driven assumptions about the environment. For example, FAB-MAP learns its likelihood models from training data [CN08]. Many visual localization systems employ learning techniques to improve the performance of its location representations, including using state-of-the-art deep learning to learn about appearance change [LGMR18] or viewpoint change [GSM18], and learning about the observation likelihoods is closely related to the chosen image representation.

The performance of the observation likelihood model can be assessed independently of the performance of the image representations. The likelihood models depend on prior belief about the likelihood of appearances which naturally vary both due to the environment itself and to the conditions under which the environment is observed.

The performance of a system can depend critically on the correctness of the likelihood model. In [Low14], a probabilistic visual localization system was evaluated on the St Lucia dataset [GMMW10]. Two observation likelihood models ($M1$ and $M2$) were trained on the data – $M1$ was trained using the data from the same time of day as the test data and $M2$ was trained using data from a different time of day. The performance of the system increased from correctly localizing in 10% of places using $M1$ to correctly localizing in over 70% of places using $M2$, with all other aspects of the system remaining unchanged.

Since the data used to train $M1$ was only captured a few hours away from the data used to train $M2$, these results suggest that not only is the correct context necessary for generating a correct location belief, but that the system must be flexible to change as the appearance of the environment varies. In fact, in some circumstances a dynamically generated likelihood model approximated online using current environment data can out-perform a pre-trained likelihood model that was calculated using exact ground truth data captured only a few hours earlier [Low14]. However, a model that is approximated online also contains assumptions about the environmental context [LM15]. If these assumptions are incorrect, it can also negatively affect the localization performance.

5 Conclusion

Visual localization has made transformative progress in recent years, with existing image description methods able to perform robustly in impressively challenging scenarios of perceptual change and differing viewpoints. These robust image description methods can be used to evaluate similarity between different location representations, which a belief generation system can convert into a likelihood or confidence metric. However, belief generation methods are sensitive to the environmental context in which the system is operating. Thus training of the models for a belief generation system must be appropriate for the current environmental conditions, and if a system is operating in a dynamic, perceptually varying environment the belief generation models must reflect this variation.

Acknowledgement

This work was supported by the Swedish Research Council (grant no. 2018-03807).

References

[AGT+18] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1437–1451, June 2018.

[BGU16] M. Brubaker, A. Geiger, and R. Urtasun. Map-based probabilistic visual self-localization. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2016.

[BHK12] H. Badino, D. Huber, and T. Kanade. Real-time topometric localization. In *2012 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1635–1642, May 2012.

[CL68] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, May 1968.

[CN08] M. Cummins and P. Newman. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.

[FB81] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981.

[GMMW10] A. Glover, W. Maddern, M. Milford, and G. Wyeth. FAB-MAP + RatSLAM: Appearance-based SLAM for multiple times of day. In *2010 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3507–3512, May 2010.

[GSM18] S. Garg, N. Sünderhauf, and M. Milford. LoST? appearance-invariant place recognition for opposite viewpoints using visual semantics. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018.

[HB14] P. Hansen and B. Browning. Visual place recognition using HMM sequence matching. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4549–4555, Sept 2014.

[KEC+18] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier. A Holistic Visual Place Recognition Approach using Lightweight CNNs for Severe ViewPoint and Appearance Changes. *arXiv e-prints*, Nov 2018.

[LBH15] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, May 2015.

[LGMR18] Y. Latif, R. Garg, M. Milford, and I. Reid. Addressing challenging place recognition tasks using generative adversarial networks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2349–2355, May 2018.

[LM15] S. Lowry and M. Milford. Building beliefs: Unsupervised generation of observation likelihoods for probabilistic localization in changing environments. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3071–3078, Sept 2015.

[Low14] S. Lowry. *Visual place recognition for persistent robot navigation in changing environments*. PhD thesis, Queensland University of Technology, 2014.

[LSN+16] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, Feb 2016.

[MW12] M. Milford and G. Wyeth. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *2012 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1643–1649, May 2012.

[NG12] T. Nicosevici and R. Garcia. Automatic visual bag-of-words for online robot navigation and mapping. *IEEE Transactions on Robotics*, 28(4):886–898, Aug 2012.

[NSBS14] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss. Robust visual robot localization across seasons using network flows. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2014.

[Ols09] E. Olson. Recognizing places using spectrally clustered local matches. *Robotics and Autonomous Systems*, 57(12):1157–1172, 2009.

[SBS07] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. pages 1–7, June 2007.

[SMT+18] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla. Benchmarking 6DOF outdoor visual localization in changing conditions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[SSD+15] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford. On the performance of ConvNet features for place recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4297–4304, Sept 2015.