COMPUTER VISION FOR ASSISTIVE INDOOR LOCALIZATION

Daniel Asmar, Adel Fakih, John Zelek
American University of Beirut, U. of Waterloo, U. of Waterloo

APPPLICATION

The number of elderly people in the world is increasing in proportion to the total number of people. As people get older, their cognitive and perceptual faculties decline in functionality and sometimes fail entirely; for example, manifesting themselves in disease or illness such as blindness or Alzheimer's disease. Approximately 14 million people in North America are affected by blindness and 6 million people suffer from Alzheimer's. Technology can help alleviate issues of confinement, security and safety as well as empowering people who feel constrained by their condition. One such way is to ensure that they are still able to conduct their daily activities by providing them with technological navigational tools. A GPS can tell people where they are and how to get somewhere but the interface can be complicated for the elderly. What if the interface was intuitively conveyed by touch like a sixth sense? This is what we have done by creating a tactile belt that we are now commercializing [TAC 11]. However, GPS does not work indoors but computer vision can be used to localize and create maps and augment as well as override GPS so navigational capabilities are never compromised (e.g., in indoor environments & urban canyons where GPS does not work).

TECHNOLOGY

Mobile smart phones provide users with a highly portable & connected computing sensor device. Their extensive sensor suite, which includes a Global Positioning System (GPS), cameras, inertial sensors, etc., demonstrates the small & portable packaging potential of a mobile connected hardware ensemble. A similar (with or without phone capability) embedded suite can be used not only for augmenting navigation for smart phone users but for other applications such as assistive devices for people with Alzheimer's and people who are blind as well as smart cognitive automobiles and augmented reality. One of the under-utilized sensors on a smart phone is the camera. Computer vision can enhance landmark-based navigation by recognizing environments, contexts and objects as well as localizing the sensor and the environment. Computer vision (i.e., image understanding) involves understanding the 3D scene creating the image. Computer vision is challenging because it is the computer that decides how to act based on an understanding of the image. Key image understanding tasks include depth computation, as well as object detection, localization, recognition and tracking. Current state of the art techniques are not able to perform any of these tasks robustly with the precision and accuracy demanded by many real-world applications unless environmental simplifications are introduced. Additional complications include operational and environmental factors. For humans, visual recognition is fast and accurate, yet robust against occlusion, clutter, viewpoint variations, and changes in lighting conditions. Moreover, learning new categories requires minimal supervision and a very small set of exemplars. Achieving this level of performance in a wearable portable system would enable a great number of useful applications including assistive devices, cognitive vehicles as well as others such as intelligent sensor augmented soldiers, real time health care and rehabilitation, etc. What we are really interested in is a sensor suite that is able to discern the state (e.g., location, activity, body position) of an agent (i.e., an agent can refer to a person or a moving vehicle) as well as the state of the agent's environment (e.g., mapping, labeling neighboring static and dynamic agents and structures). Another component of this system's architecture is the

user interface to the agent. Our interests are in wearable and tactile interfaces. These haptic interfaces can act as a substitute for another sensor (i.e., in the case of an assistive device) or to draw the attention of the agent towards or away from a location or situation.

STATE OF THE ART

Global localization of an agent outdoors is possible via GPS. In some environments such as indoors, forests, urban canyons, sub-sea, extra-terrestrial, inside mines, GPS information is not readily available. In these environments, it is still possible for an agent to localize itself by a process in the robotics field known as Simultaneous Localization and Mapping (SLAM). Via its sensor suite, in SLAM, an agent builds a map of the environment while simultaneously localizing itself with respect to that map. If a map is available a priori, navigation is relatively simple. If the global position of the agent is known at every time (i.e., using GPS outdoors), mapping of the environment is considered a solved problem [BUR 99]. The problem becomes complicated when the pose of the agent and the location of the landmarks are not available. This is because the errors in the pose of the robot and landmark are correlated. SLAM manages error growth by maintaining a covariance matrix, which correlates all the errors of the state vector including the position and bearing of the agent as well as all landmarks. At each time step, the covariance matrix is propagated forward in time, indicating the interdependence between variables. The Kalman Filter (KF), Extended Kalman Filter (EKF), or the Particle Filter (PF) are the conventional mechanisms by which the state and co-variances are propagated between time steps. The SLAM process is hard because of (1) dimensionality explosion; (2) the need for robust landmark detection; and (3) data association issues. Since the seminal work in SLAM [SMI 86], recent developments have strengthened SLAM theory [DUR 06a]. This development was aided by only using simple feature landmarks, which can be easily detected and recognized. Moving the SLAM method into more challenging environments necessitates that the landmark issue be addressed. The most common sensor used is a range-bearing sensor: sonar for indoor and underwater and the Laser Range Finder (LRF) for outdoor situations. However, LRFs are only 1D sensors and similar landmarks are impossible to differentiate. A camera is a richer alternative sensor that provides not only depth but also color and texture. Most of the Vision SLAM systems implemented to date use saliencies or Interest Points (IP) [NEI 08]. Some of the challenges for vision SLAM include the acquisition of depth and the use of natural landmark objects to avoid data explosion. A stereo rig can be used but the precision of depth values is proportional to its baseline and inversely proportional to the square of the observed depth, which limits use in outdoor environments [JUN 04]. The ultimate goal for vision SLAM remains to be operational anywhere using the environment's natural features as landmarks with passive vision. As with other types of SLAM systems, loop closures are also a concern. A related problem to SLAM is called visual odometry in robotics, or structure-from-motion (SFM) in computer vision literature [STU 09]; essentially this is SLAM without keeping the built map around.

For SFM, depth is determined by triangulation once eliminating depth and constraining the problem space estimate of motion. The optimal solution for solving this problem is referred to as Bundle Adjustment [TRI 99], which is an off-line approach, which maximizes the likelihood of the 3D parameters given the image projections. Standard on-line approaches are (1) filter based which utilize the Markovian assumption; (2) odometry approaches [KON 07] which capitalize on matching many features in the last few frames only; and (3) key frame approaches which select the set of key frames over the whole sequence from which to base the calculations about. Filter based approaches utilize EKF, Particle Filters or variants such as a Graph of Local Filters (GLF) [EAD 07] techniques in a recursive predict and update scheme. A Key frame approach [KLE 08] optimizes only the motion over key frames and not all the frames. The best SFM approaches appear to consist of a front-end tracker, which can be filter-based, and an optimization back end such as the GLF or key frame approach. Filtering improves results especially when the overall processing budget is constrained. Challenges and issues

for on-line SFM (as well as off-line) include occlusion, dynamic environments as well as dense structure estimation.

The concept of knowing an object also embeds information about what we can do with those objects. Object perception based on appearance unfortunately does not always determine function as some objects; for example, a chair may take odd forms. Object recognition/categorization is hard and has its challenges including: (1) viewpoint variation; (2) illumination changes; (3) occlusion; (4) scale; (5) deformation; (6) background clutter; and (7) intra-class variation. From a statistical point of view, if we want to detect a car in the image, we really want to know P(car|image) vs. P(nocar|image). Applying Bayes rule:

$$\frac{P(car|image)}{P(nocar|image)} = \frac{P(image|car)}{P(image|nocar)} \frac{P(car)}{P(nocar)}$$

where P(car|image)/P(nocar|image) is a posterior ratio, P(image|car)/P(image|nocar) is the likelihood ratio and P(car)/P(nocar) is the prior. Discriminative methods directly model the posterior. Generative methods model the likelihood and priors. There are three main issues with object category recognition: (1) representation; (2) learning; and (3) recognition. The method has direct bearing on the representation. A learning phase is necessary for both types of methods. Two types of generative models are the bag-of-words approach [SIV 05] and the method based on parts-based models. The bag-of-words model chops an image into patches with indifference to the location of the patches. The patches may be based on grid cutting or using a feature detector [MIK 04]. The patches are put into a codeword dictionary, which is classified. This same dictionary is used for recognition. One of the strongest criticisms of this method is the lack of geometric and spatial information. Recent efforts for including spatial information include using correlograms [SAV 06], incorporating parts models to add constraints [SUD 07] as well as using a pyramid spatial partitioning to constrain the detected features [LAZ 06]. Even with these attempts, bag-of-words methods still suffer from location problems as well as correspondence issues, which are inherent within and amongst objects. Part based approaches include time-consuming a priori geometric information [ZHU 06]. There is an explicit notion of correspondence between the image and model. Efficient methods exist for a large number of parts and positions in the image. Hierarchical models also allow for more parts [FID 09]. Discriminative (or Classifier based) methods cast the object detection and recognition problem as a classification problem [TOR 04]. The image is partitioned into a set of overlapping windows and a decision is taken at each window on whether it contains the target object or not. Other recent developments have shown the importance of context in object recognition [JIN 10] and that 3D is inherent and can actually even be computed from a still image [SAX 09]. There has been some success with 3D object category recognition [SUN 09].

ADDRESSING CURRENT LIMITATIONS

We have addressed the state-of-the-art shortcomings in all of the three components – SLAM, SFM, object recognition – necessary for performing localization and recognition using computer vision. We used tree trunks as a natural landmark to demonstrate SLAM with computer vision [ASM 09]. The results were demonstrated in a park setting, the camera triangulated between tree trunks in order to determine their position and the global location of the camera. The results were compared against GPS ground truth and the results were within 1 cm accuracy. A stereo camera was used to compute depth. SFM can be used as an alternative to compute depth. A collection of real time filters were developed [FAK 09] and showed that we could compute depth with a single camera on the fly. The results were just as accurate as batch processing of the video sequence. Batch processing is an off line process where the video data is first collected and then processed which cannot be used for online processing. Rather than just focusing on using a single landmark, we developed an object class recognition algorithm [FAZ 07] that takes the best of all the current methods. We were able to achieve recognition rates close to 100% accuracy for standard benchmark databases.

The computer vision algorithms that we have developed (i.e., SLAM, SFM, object recognition) are the building blocks for

localization in environments where no GPS is available. We now plan to fuse our efforts into a single process for that very purpose. Another hurdle is to package these algorithms so that they execute on a low power microprocessor, as a real-time, wearable, portable solution is desirable. Besides tracking people with Alzheimer's and people who are blind in home settings, this technology can also be used in nursing homes and for applications where navigational assistance is necessary when cognitive and perceptual capabilities are compromised.

REFERENCES

[1] [BUR 99] BURGARD W., CREMERS A., FOX D., AHNEL D., LAKEMEYER G., D. SCHULZ, STEINER W., THRUN S., « Experiences with an interactive museum tour-guide robot », Artificial Intelligence, vol. 114, no 1-2, 1999, p. 3-55.

[2] [DUR 96] DURRANT-WHYTE H., « An autonomous guided vehicle for cargo handling applications», Int. Journal of Robotics Research, vol. 15, no 5, 1996, p. 407-441.

[3] [DUR 06a] DURRANT-WHYTE H., BAILEY T., « Simultaneous Localization and Mapping: Part 1 », IEEE Robotics and Automation Magazine, , 2006, p. 99-108.

[4] [DUR 06b] DURRANT-WHYTE H., BAILEY T., « Simultaneous Localization and Mapping: Part 2 », IEEE Robotics and Automation Magazine, , 2006, p. 108-117.

[5] [EAD 07] EADE E., DRUMMOND T., « Monocular slam as a graph of coalesced observations», Proc. 11th IEEE Int. Conference on Computer Vision, 2007.

[6] [ELF 89] ELFES A., « Using occupancy grids for mobile robot perception and navigation », Computer, vol. 22, no 6, 1989, p. 46-57.

[7] [FID 09] FIDLER S., BOBEN M., LEONARDIS A., « Learning Hierarchical Compositional Representations of Object Structure », DICKINSON S., LEONARDIS A., SCHIELE B., TARR M., Eds., Object Categorization : Computer and Human Vision Perspectives, Cambridge University Press, 2009.

[8] [JIN 10] JIN-CHOI M., LIM J., TORRALBA A., WILLSKY A. S., « Exploiting Hierarchical Context on a Large Database of Object Categories », Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2010.

[9] [JUN 04] JUNG I.-K., « Simultaneous localization and mapping in 3D environments with stereo vision », PhD thesis, CNRS, Toulouse, France, 2004.

[10] [KLE 08] KLEIN G., MURRAY D., « Improving the agility of keyframe based slam », ECCV, p. 802-815, 2008.

[11] [KON 07] KONOLIGE K., AGRAWAL M., SOLA J., « Large-scale visual odometry for rough terrain », Proc. Int. Symposium on Robotics Research, 2007.

[12] [LAZ 06] LAZEBNIK S., SCHMID C., PONCE J., « Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories », Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, p. 2169-2178, IEEE, New York, June 2006.

[13] [MIK 04] MIKOLAJCZYK K., SCHMID C., « Scale & Affine Invariant Interest Point Detectors », International Journal of Computer Vision, vol. 60, p. 63-84, Kluwer Academic, 2004.

[14] [NEI 08] NEIRA J., DAVISON A., LEONARD J., « Guest Editorial Special Issue on Visual SLAM », IEEE Transactions on Robotics, vol. 24, no 5, 2008, p. 929-931.

[15] [SAV 06] SAVARESE S., WINN J., CRIMINISI A., « Discriminative Object Class Models of Appearance and Shape by Correlations », Proc. IEEE Computer Vision and Pattern Recognition (CVPR), IEEE, 2006.

[16] [SAX 09] SAXENA A., SUN M., NG A. Y., « Make3d : Learning 3D Scene Structure from a single still image », IEEE Transactions of Pattern Analysis and Machine Intelligence, vol. 30, no 5, 2009, p. 824-840.A. First, B. Second, and C. Third, "Example of a reference," *J. Examples*, vol. 34, pp. 34-49, 2003. (*References*)

[17] [SIV 05] SIVIC J., RUSSELL B. C., EFROS A. A., ZISSERMAN A., FREEMAN W. T., « Discovering Object Categories in Image Collections », Proceedings of the International Conference on Computer Vision, 2005.

[18] [SMI 86] SMITH R., CHEESEMAN P., « On the representation and estimation of spatial uncertainty », Int. Journal of Robotics Research, vol. 5, no 4, 1986, p. 56-68.

[19] [STU 09] STURM J., VISSER A., « An apearance-based visual compass for mobile robots », Robotics and Autonomous Systems, vol. 57, no 5, 2009, p. 536-545.

[20] [SUD 07] SUDDERTH E. B., TORRALBA A., FREEMAN W. T., WILLSKY A. S., « Describing Visual Scenes using Transformed Objects and Parts », International Journal of Computer Vision, vol. 77, 2007.

[21] [SUN 09] SUN M., SU H., SAVARESE S., FEI-FEI L., « A Multi-view probabilistic model for 3D Object Classes », Computer Vision and Pattern Recognition, 2009.

[22] [TOR 04] TORRALBA A., MURPHY K., FREEMAN W., « Sharing features : efficient boosting procedures for multiclass object detection », Proc. of the IEEE Computer Socity Conference on Computer Vision and Pattern Recognition, 2004.

[23] [TRI 99] TRIGGS B., MCLAUCHLAND P., HARTLEY R., FITZGIBBON A., « Bundle adjustment - a modern synthesis », ICCV 99 Proceedings of the International Workshop on Vision Algorithms, p. 298-372, Springer Verlag, 1999.

[24] [ZHU 06] ZHU S. C., MUMFORD D., A Stochastic Grammar of Images, Foundations and Trends in Computer Graphics and Vision, 2006.

[25] [TAC 11] Tactile Sight Inc., www.tactilesight.com, 2011.

[26] [ASM 09] Daniel C. Asmar and John S. Zelek and Samer M. Abdallah, (2009), Vision SLAM Maps: Towards Richer Content, in Design and Control of Intelligent Robotic Systems, Series: Studies in Computational Intelligence: Vol. 177; editors: Dikkai Liu, Lingfeng Wang, Kay Chen Tan.

[27] [FAK 09] Adel Fakih and Samantha Ng and John S. Zelek, (2009), Improving GPS Localization with Vision and Inertial Sensing, Geomatica, 63(4), pp 139-146.

[28] [FAZ 07] Ehsan Ersi Fazl and J.S. Zelek, (2007). "Local Graph Matching for Object Category Recognition", CRV07, Montreal, QC, 2007.