Visual Nouns for Indoor/Outdoor Navigation

Edgardo Molina[1], Zhigang Zhu[1], and Yingli Tian[2]

[1] Department of Computer Science
[2] Department of Electrical Engineering,
Grove School of Engineering, The City College of New York,
138th Street and Convent Avenue, New York, NY 10031
{molina,zhu}@cs.ccny.cuny.edu
ytian@ccny.cuny.edu

Abstract. We propose a local orientation and navigation framework based on visual features that provide location recognition, context augmentation, and viewer localization information to a human user. Mosaics are used to map local areas to ease user navigation through streets and hallways, by providing a wider field of view (FOV) and the inclusion of more decisive features. Within the mosaics, we extract "visual noun" features. We consider 3 types of visual noun features: signage, visual-text, and visual-icons that we propose as a low-cost method for augmenting environments.

1 Introduction: Idea and Impact

Local indoor and outdoor navigation and localization remains a challenging problem. Various solutions have been proposed with varying degrees of success. GPS and GPS combined with image registration works well outdoors and for large area localization, but can be problematic in dense urban environments and indoors since devices require a direct view of the sky. Augmented indoor positioning systems have been proposed [8] using RFID or sonar sensors. Such systems require extensive and expensive environment augmentation, and can suffer from interference in noisy (from both radio-frequencies and acoustic) environments and power restrictions. A vast amount of research has focused on robot navigation and SLAM (Simultaneous Localization And Mapping). A smaller subset of work has focused on adapting the research to human users, in particular users that are blind or low-vision.

Here we propose a local orientation and navigation framework based on visual features that provide location recognition, context augmentation, and viewer localization information to a human user. Although it seems counter-intuitive to use visual features for blind and low-vision user navigation, we note that signs, icons, and text in images are among the most common ways of providing humans context information. The key is being able to perform object-recognition and text recognition from video reliably so that it can be communicated to a blind or low-vision user with text-to-speech software. Furthermore, these features in the scene could also provide the user accurate location information in the 3D world. If we consider the image features traditionally used in robotics for localization: image edges, corners, SIFT/SURF

K. Miesenberger et al. (Eds.): ICCHP 2012, Part II, LNCS 7383, pp. 33–40, 2012.

descriptors and so on, we realize that while they work well in algorithms it would have almost zero benefit to communicate such information to a human user.

In Section 2 we review related work. In Section 3 we fully describe what visual nouns are. Section 4 presents our visual noun based algorithms. Section 5 shows experiments and results. Conclusions and a discussion of further work are in Section 6.

2 Related Work

A lot of work has been done in object detection and recognition. For object detection a useful method has been the MSER blob detector [14]; it has been extended to handle color [12], and text [3]. Saliency maps are another method employed in detecting objects and areas of interest [1]. Both MSER and Saliency map methods provide regions of interest that are consistent with characteristics we expect in visual features, such as signs and text, mainly because they highly contrast with their backgrounds. Object matching has been well studied, with simple methods such as template matching, to machine learning based methods. Object detections can often be distinguished from one another readily, but to be truly informative to human users we must recognize the sign from a labeled database to communicate its meaning to the user.

We use the visual features to perform localization of the user in their environment. Methods such as 3D reconstruction [16] can be employed, or methods with a sparse set of features can also be used, such as the PnP algorithm [15]. The PnP algorithm requires some knowledge or mapping of the signs in 3D space.

The typical camera view is not wide enough to cover enough visual features for a user to perform localization. A sighted user usually looks around to find recognizable features around them. Similarly the blind and low-vision do the same to get an understanding of sounds around them. Using a panorama of a user's surroundings provides more visual features that can help localize the user. Visual navigation using panoramic images have been studied by us [6] and others [5], and here we leverage our past experience and integrate visual nouns as local features for user localization using panoramic images.

The system presented here differs from the typical SLAM approaches in that we are not interested in automatically mapping entire scenes, but rather providing salient local orientation and localization information to a human user, who is using their own cognitive abilities to make decisions.

3 Visual Nouns in Context: Our Approach

A primary goal in this work is to use and detect features that naturally provide human users context information, not only what they see, but also as to where they are in the 3D world. Below we describe the 3 types of features we call Visual Nouns:

Text appearance is a rarely used feature in video and image matching and retrieval applications. Traditionally, OCR algorithms reduce and map text in imagery to ASCII character codes, occasionally with some minimal formatting/layout information. Visually, text provides richer features such as: font styling, color/texture, its geometric

alignment, and size relative to other text. In addition, each text sign may contain unique markers due to age, weathering, damage, and vandalism. In outdoor and indoor navigation scenarios, users encounter such text on storefronts, signs, postings, and doors. Since, typically these are static and on planar surfaces, we may wish to use visual text as fiducial features for localization. Recent work [2, 3] has presented methods for extracting visual text for better performance in information retrieval and matching. The results in these works provide motivation to further extend the work in particular for 3D localization, when building visual navigation systems for the blind.

In addition to aiding navigation, combining Text with other signage provides users (the blind especially) location context information. When combined with a text-to-speech component, blind users can be alerted as they approach and arrive at known locations, such as health facilities, restaurants or friends and family homes.

Visual-icons denote universal symbols that are in use throughout the world that convey a particular meaning. The Department of Transportation in the US has a set of vehicle and pedestrian symbols that are similar to those used in other countries to depict where a user can find a train, taxi, elevators, escalators, etc. Figure 1 shows 5 sample icons. Such symbols are not universally standardized, but there are efforts to create databases of such symbols [9,10].

Fig. 1. Five Aiga & US DOT symbols from [9]

Augmenting an environment with electronic positioning devices (RFID, NFC) is always a costly endeavor. Using symbols is more cost effective since they can be printed and only requires cameras for detection, which are already widely available. Additionally, these signs can be further augmented as the price of electronic tags and receivers fall.

Signage as used in our paper refers to those signs that are not already covered by Text or Visual Icons. In general signs are natural for matching as they are found both indoors and outdoors. Signs contain logos, text, and symbols that in addition to serving as localization markers also provide contextual information. These especially become useful in recognition and verbal translation for the blind and visually impaired. Here we differentiate visual-icons as those we are matching against a known database of universal symbols, and we restrict it to binary image symbols. With signage we refer more generally to all signs (grayscale and colored), including previously unseen signage (not in a DB of symbols) and logos and brand marks, such as a pizza image outside of a pizzeria or car brand mark at a car dealership.

4 Visual Noun Based Localization: Algorithms

We propose the use of Visual Noun features to aide blind and low-vision users in orientation and navigation tasks. Our system considers a user with a wearable camera (either on the frame of glasses or on a cap). They arrive at a place that is new to them,

has recently changed or they have not immediately recognized. The user may survey the area around them by panning their head; blind and low-vision users may also do this as they try to locate distinct sounds or lighting in their surroundings. Our system:

1. captures video of the panned area;
2. registers video image frames to the first frame (will serve as reference);
3. generates a wide field-of-view panorama (see Figures 3 and 4);
4. extracts and matches visual-noun features; and
5. localizes the user in 3D space relative to the visual nouns.

Figure 2 shows a workflow of the entire process. The output available to the user is the detected visual nouns and their meaning through matched visual-icons and text in the surrounding panorama. Further, the user's location relative to the visual nouns can be estimated using the PnP algorithm. The final step is using an interface and text-to-speech so that the user can utilize the discovered information, here we have made no attempt for this final step but we plan to explore it further.

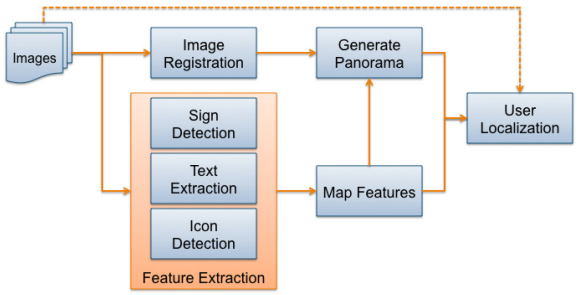

Fig. 2. Visual-noun based navigation framework

4.1 Mapping

To assist users in navigation tasks we must first obtain or build a mapping of the environment they are navigating in. In the US, Google has the StreetView service and Microsoft has a similar service, which they both continue to grow. Google's Street-View provide panoramic views of major city streets. Microsoft's Bing maps provide route-panoramas like the ones in [11] of many major city streets. For indoor and outdoor environments where no pre-existing mappings exist, local mosaic maps can be built using various existing techniques [7,11] which all produce aligned panoramas. In the City College Visual Computing Lab, we have developed software packages to generate panoramic mosaics in real-time with a hand-held camera [7]. Visual nouns in both the conventional images and panoramic images are be annotated with their 3D location information and therefore these visual nouns and their images can be used as the scene model for both recognition and localization.

4.2 Visual Nouns Extraction

Visual-noun features are extracted from images by locating areas where high contrast changes occur. The intuition being that signage, text, and icons used to alert users should have enough contrast so that they catch the viewer's attention. In this work we have used the MSER blob detector [13] which performs well at segmenting regions with high contrast from their backgrounds, as proposed by Chen et. al. [3]. It was found that MSER does not deal well with blurry regions and so the Edge Enhanced Maximally Stable Extremal Regions (EE-MSER) algorithm was proposed in [3] to detect text in natural images. We found a sharpening procedure corrected some of the issues that occur due to motion blur in video.

4.3 Localization

By combining multiple visual nouns with known 3D locations, the viewer's pose can be determined, using a typical pose estimation algorithm such as the PnP algorithm [4] we have used in multi-robots navigation tasks.

A panorama view allows us to match both reliable visual-nouns and traditional features that are common among multiple views. It may be the case that the visible visual-nouns are not enough to perform localization using the PnP algorithm. In these cases we augment the visual-nouns with traditional features that are consistent with the panoramic view and with the projections of the visual-nouns. We use RANSAC to check that all features provide consistent projections onto the panorama.

With the panoramic view we can take 3 algorithmic approaches to assist blind users.

1. Through visual noun detection we can provide the user contextual information about their surroundings. (User can be told what resources/facilities are around them.)
2. The panoramic view provides the user orientation information, which can be used to tell a user in which direction they should turn. (User can be told which sign they are facing.)
3. When the visual noun locations are known and panoramas constructed we can localize the user in 3D space using the PnP algorithm. (User can be told how far from a sign they are.)

5 Experiments and Results

In our experiment we augmented an indoor hallway with 4 visual-icon printouts and we captured a 5 second video recording of the surroundings. Figure 3 shows 3 original video frames. Few visual nouns are often seen in any single view. Figure 4 shows the result from registering the video frames and generating a wide field-of-view panorama of the scene. The panorama contains many more visual-noun features which provide both context and allow us to localize a user.

Fig. 3. Three original frames from the 5 seconds of video

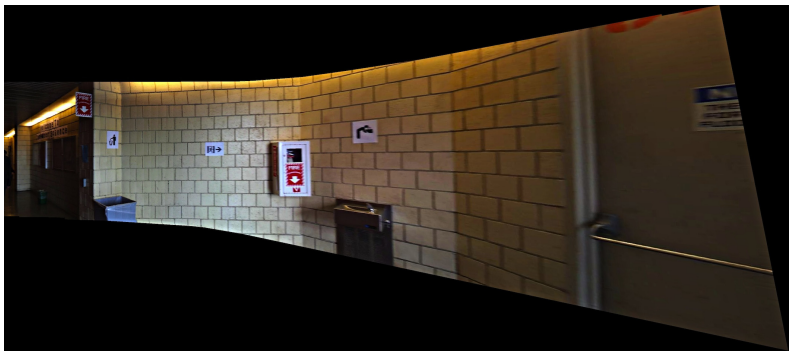

Fig. 4. Wide field-of-view panorama generated from video

Using the panorama we are able to identify signs and text markings across the entire scene. We then use MSER to detect regions of high contrast. Figure 5 shows the resulting detections on various signs. Some false positives are also detected by the MSER algorithm, but can be reduced by matching against a database of visual-icons, and applying geometric text filters as described in [3].

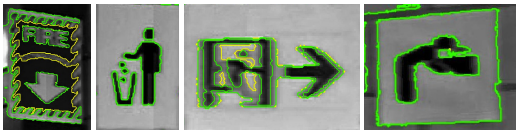

Fig. 5. MSER results on signs

Figure 6 shows a single view and a panorama with markings to the right of the features used in the localization experiment. The red circle markings denote visual nouns, and the green square markings denote additional selected points.

Table 1 shows the results from our localization experiment. The test video was captured with a hand held camera by a viewer at head height. For the single frame, the *Estimated Pose* row gives our manually measured camera pose. The *Single View* row shows the results of using the 4 marked features, showing that we were up to 3 feet off in any one axis, and a few degrees off from the estimate. Using the panorama (with the same reference frame as the single view) we located 8 features which gave us a better overall pose estimate. Using the 8 features from the panorama, our translation result was only 6 inches off from the estimated pose (along the Y axis, the distance to the wall).

Fig. 6. A single view and a panorama marked with detected visual-nouns in red circles, and additional features in green squares (markings are to the right of features)

Table 1. Visual-Noun localization experiment results

	(Tx, Ty, Tz) inches	(θx, θy, θz) degrees
Estimated Pose	(-30, 140, 65) manual est.	(-90, 20, 180) manual est.
Single View	(3.043, 136.23, 73.39)	(-98.07, 6.05, 179.79)
Panorama	(-29.90, 143.30, 66.29)	(-94.28, 17.59, -179.04)

The green square markings denote feature points which were manually measured and used in localization. With a single view, often insufficient features are detected to robustly run localization. The panorama provides us more features, but we are exploring the use of matched SIFT points as secondary features to the Visual-Nouns to make localization more robust.

6 Conclusions and Discussions

In this paper we have proposed the use of Visual Noun features as a way of providing users, particularly the blind and low-vision, context information that is otherwise difficult for them to obtain. Visual-noun features such as signage, visual-text, and visual-icons, are already found throughout human indoor and outdoor environments. And with the availability of inexpensive cameras and improved vision algorithms, we can readily detect them without significant infrastructure investment. We present a workflow that adapts core robotics and vision algorithms for use with a human user that will be using their own cognitive skill in making navigation decisions. We have presented preliminary experiments and results that demonstrate the capabilities of using Visual Nouns for navigation and localization tasks.

There are many open questions we are still exploring in making this system useful to a user. The first is the interface the user would interact with. We recognize that this is crucial to successfully aide users and we have focused in this paper on ensuring that the contextual and 3D location information available is in a human understandable format.

Acknowledgments. This work is supported by US National Science Foundation Emerging Frontiers in Research and Innovation Program under Award No. EFRI-1137172, and City SEEDs: City College 2011 President Grant for Interdisciplinary Scientific Research Collaborations.

References

1. Wang, S., Tian, Y.: Indoor Signage Detection Based on Saliency Map and Bipartite Graph Matching. In: Intl. Workshop on Biomedical and Health Informatics, BHI (2011)
2. Schroth, G., Hilsenbeck, S., Huitl, R., Schweiger, F., Steinbach, E.: Exploiting Text-related Features for Content-based Image Retrieval. In: IEEE Intl. Symposium on Multimedia (ISM), Dana Point, CA, USA (December 2011)
3. Chen, H., Tsai, S.S., Schroth, G., Chen, D.M., Grzeszczuk, R., Girod, B.: Robust Text Detection in Natural Images with Edge-enhanced Maximally Stable Extremal Regions. In: 2011 IEEE Intl. Conference on Image Processing, Brussels (September 2011)
4. Feng, Y., Zhu, Z., Xiao, J.: Self-Localization of a Heterogeneous Multi-Robot Team in Constrained 3D Space. In: IEEE/RSJ Intl. Conference on Intelligent Robots and Systems, San Diego, CA, USA, October 29-November 2 (2007)
5. Binding, D., Labrosse, F.: Visual Local Navigation Using Warped Panoramic Images. In: Proceedings of Towards Autonomous Robotic Systems, Guildford, UK, pp. 19–26 (2006)
6. Zhu, Z., Karuppiah, D.R., Riseman, E.M., Hanson, A.R.: Keep Smart, Omnidirectional Eyes on You - Adaptive Panoramic Stereo Vision for Human Tracking with Cooperative Mobile Robots. Robotics and Automation Magazine, Special Issue on Panoramic Robots 14(11), 69–78 (2004)
7. Zhu, Xu, G., Riseman, E., Hanson, A.: Fast Construction of Dynamic and Multi-Resolution 360° Panoramas from Video Sequences. Image & Vision Computing Journal 24(1), 13–26 (2006)
8. Zhu, Z., Huang, T.S. (eds.): Multimodal Surveillance: Sensors, Algorithms and Systems. Artech House Publisher (July 2007)
9. Symbol-Signs, `http://www.aiga.org/symbol-signs/`
10. The Noun Project, `http://www.thenounproject.org/`
11. Zheng, J.Y.: Digital Route Panoramas. IEEE Multimedia 10, part 3, 57–67 (2003)
12. Forssen, P.-E.: Maximally Stable Colour Regions for Recognition and Matching. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, June 17-22, pp. 1–8 (2007)
13. Vedaldi, A., Fulkerson, B.: VLFeat, An Open and Portable Library of Computer Vision Algorithms (2008), `http://www.vlfeat.org`
14. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proc. of British Machine Vision Conference, pp. 384–396 (2002)
15. Quan, L., Lan, Z.: Linear N-point camera pose determination. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(8), 774–780 (1999)
16. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2003)