Generating a Map from a Linguistic Description

ECE 4995: Undergraduate Honors Research
May 11, 2009

Andrew Buck, Undergraduate Researcher

Isaac Sledge, Graduate Researcher
James Keller & Marjorie Skubic, Faculty Mentors

Contents

Abstract

Humans have an innate reasoning ability that allows for the conversion of a verbal description into a real world location. Computers can mimic this process by breaking it into several smaller problems. These include speech recognition, deep-language understanding, spatial reasoning, and geospatial image matching. Although researchers have explored each of these fields extensively, they have not yet combined them into a complete system. In this paper, we explore the possibilities and limitations of such an automated system with a focus on spatial reasoning and geospatial image matching.

Introduction

Imagine a person walking along a street describing his location to another person using only his voice. There would be a number of steps involved with conveying the information. First, the listener must form words out of the sound waves making up his voice. Then, he must interpret these words into a language and group them such that they carry some useful information about the scene. At this point, the listener has constructed a linguistic tree that expresses all of the objects in the scene and their relations to one another. Now, the listener begins to place each of these objects into a mental "sketch" using the spatial relationships between objects. If the listener is familiar with the area, he may be able to match his sketch to a real world location and complete the process. Figure 1 illustrates each step of the conversion process from a verbal description to a real world location.

Figure 1: An illustration of the conversion process from a verbal description to a real world location.

The steps that a human mind must go through to match a description to a location are very similar to what a computer must do to achieve the same result. The first half of the problem involves speech recognition and a deep understanding of the English language (assuming that the original description was in English). The second half of the problem uses several image-processing techniques, many that come from the field of geospatial intelligence. The techniques themselves are only as good as the data they work with, so a successful translation requires both good input and reference data. In this paper, we investigate primarily the second half of the process, beginning with the derivation of spatial relations from linguistic descriptions and the placement of objects into a sketch. Then we look at methods for segmenting satellite images, a key requirement to matching a sketch to a real location. Lastly, we explore the requirements of the linguistic description in an example sketch.

Text to Sketch

For the Text to Sketch module of the process, we begin with a linguistic tree like the one in Figure 2. Each sentence of the description maps into one of these trees. From these trees, we can extract much information regarding the structure of the sentence. Eventually it is possible to build a small dataset that contains each physical object and their relationships with one another. For example, consider the sentence "To my immediate left, I see a small building surrounded by a parking lot [1]." From this sentence, we can tell that there are two objects and an actor, denoted by the word "I." Both objects are to the left of the actor and one object surrounds the other. The phrase "immediate left" shows some of the ambiguity inherently present when describing spatial relationships using linguistic descriptions. To manage this ambiguity, we turn to the histograms of forces.

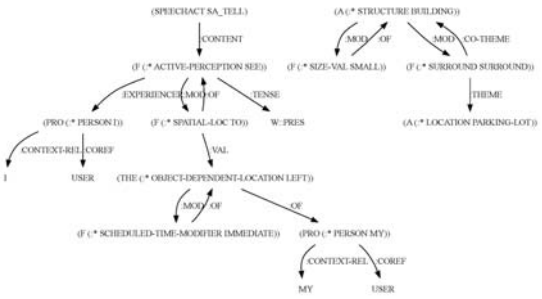


Figure 2: Logical Form Graph produced from the sentence "To my immediate left, I see a small building surrounded by a parking lot."

Histograms of Forces

Given a pair of crisp, two-dimensional objects A and B, we want to know the amount that A is in direction θ from B [1]. We define a function $F_r^{AB}(\theta)$ that evaluates the amount of support for which this is true. Given a line $\Delta_\theta(v)$ as in Figure 3, we can determine the set of line segments that intersect objects A and B. We evaluate each segment in A against each segment in B. For each point in the first segment, we measure the distance d_{MN} to a point in the second segment. We can then define functions that process points: $\phi_r(M-N) = 1/d_{MN}^r$, line segments: $f_r\left(d_I, d_{IJ}^\theta, d_J\right) = \int_{a_I^\theta}^{b_I^\theta} \int_{a_J^\theta}^{b_J^\theta} \phi_r(u-v)dv\,du$, longitudinal sections: $\mathcal{F}_r\left(\theta, A_\theta(v), B_\theta(v)\right) = \sum_{i,j} f_r\left(d_{I_i}, d_{I_iJ_j}^\theta, d_{J_j}\right)$, and directions: $F_r^{AB}(\theta) = \int_{-\infty}^{\infty} \mathcal{F}_r\left(\theta, A_\theta(v), B_\theta(v)\right) dv$. We can evaluate these functions with different values of r to capture different information. For example, $r = 0$ gives the constant force, which represents object independence from distance, and $r = 2$ gives the gravitational force, which represents object independence from scale. Plotting this function for all values of θ gives the histograms of forces. These histograms represent the relative position of the two objects in a non-specific way that translates well into English generalities.

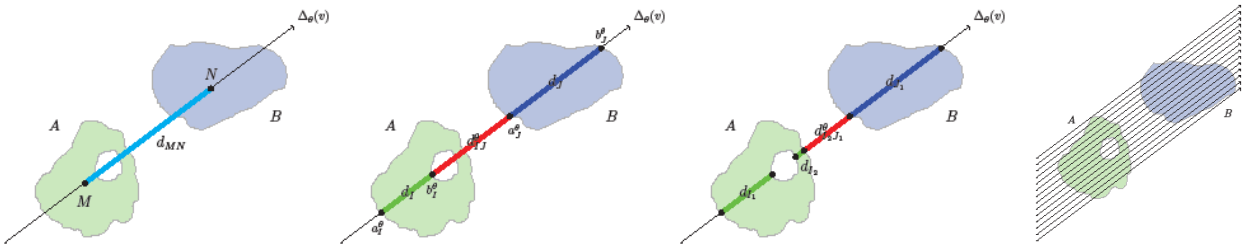


Figure 3: Overview of the histograms of forces computation for two crisp objects.

In the example shown in Figure 3, we might describe object B as being "to the front and right" of object A. Building the histograms of forces, we see from Figure 4 that the majority of the plot lies between the angles 0 and $\frac{\pi}{2}$. If instead of the object positions, we knew the phrase "to the front and right," we could perform this process in reverse. We could construct the histograms of forces and use this information to derive the object positions. This is the technique used to place objects into a sketch, known as a fuzzy region template.

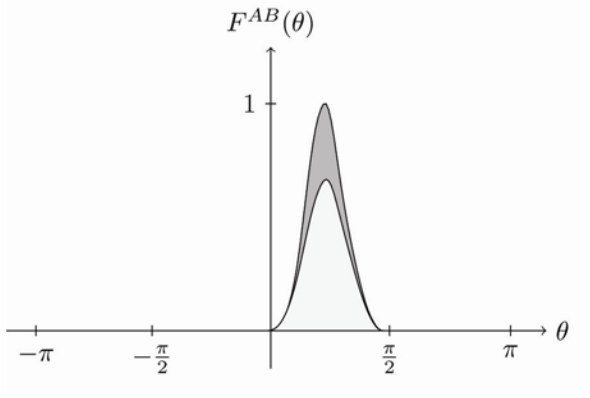


Figure 4: The histograms of constant and gravitational forces for the objects in Figure 3.

Fuzzy Region Templates

In order to place an object into a sketch, we must know its direction and distance from an object already in the sketch. This information comes from the linguistic description and we can use it to build the histograms of forces. By reversing the previous process, we can determine the spatial relationship between two objects and develop a mask that constrains the area where we should place the new object [1]. For example, assume that we have the sentence "The columns are in front of and somewhat close to Jesse Hall." If we know the position of the base object, Jesse Hall, then we can say "in front of" is in the direction $\frac{\pi}{2}$, or directly above Jesse Hall. Adding a distance requirement further limits the available placement area. The distance "somewhat close" is rather ambiguous and requires that we make some interpretation such as in Figure 5.

Figure 5: The process of building a fuzzy region template for placing the University of Missouri columns in front of Jesse Hall.

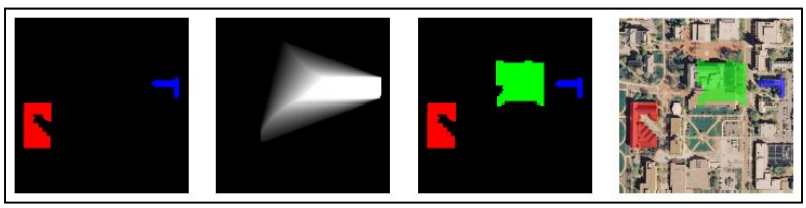


Figure 6: Adding an object with two reference objects.

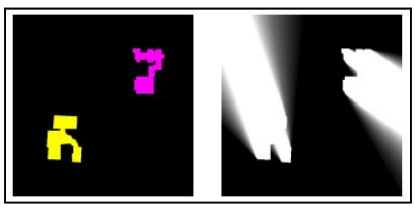


Figure 7: A multi-part description that cannot be directly resolved.

If we have enough information, it may be possible to use multiple reference points to place an object, as in Figure 6. Here, we describe the green building as being "perfectly to the left of the blue building" and "mostly to the right, but somewhat above the red building." The resulting search area is in white between the red and blue buildings. An issue arises when the search areas do not intersect as in Figure 7. This can occur because of a poor description by the human or an incorrect translation by the computer system. In this case, we must modify the search areas until they intersect or remove one of them completely. If we have enough accurate information, it is possible to use this process of fuzzy region templates to build a rudimentary sketch of an area. Refining this sketch and matching it to a real world location requires that we also have accurate satellite image data.

Image Segmentation

Satellite imagery has become available in recent years with a high enough spatial resolution to be useful in urban environments. Resolutions of less than one meter are now available to the public with the latest commercial satellites. There is a wealth of information in these images, and it is constantly changing due to new building construction and demolition. Identifying objects in these images is one of the key requirements for a complete Text to Sketch system. Image segmentation is the process of locating and labeling objects in an image either automatically or by hand. Automatic segmentation offers many benefits such as quick updates of changing areas, however the accuracy still lags far behind hand segmentation.

There are many different techniques for segmenting an image automatically. Each method has strengths and weaknesses and the best results often come from combining several different methods. For example, methods that use spatial filtering are well suited to classify manmade structures while spectral based methods tend to be better suited for classifying vegetation and soils [2]. Among the best methods for identifying buildings and structures are those that generate object-based hierarchies. In these methods, we assign individual pixels an initial classification and then group them together into progressively larger clusters [3]. These clusters map into a multilevel hierarchy based on size, which we can then use to classify objects. Figure 8 shows three different cluster sizes, which make up the roof of a building and help to classify similar pixel clusters.

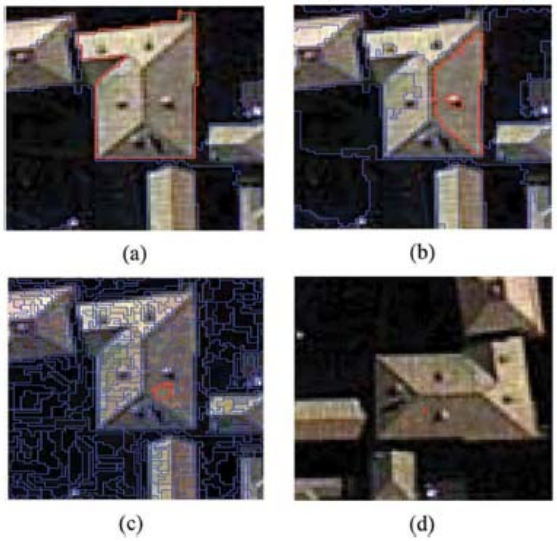
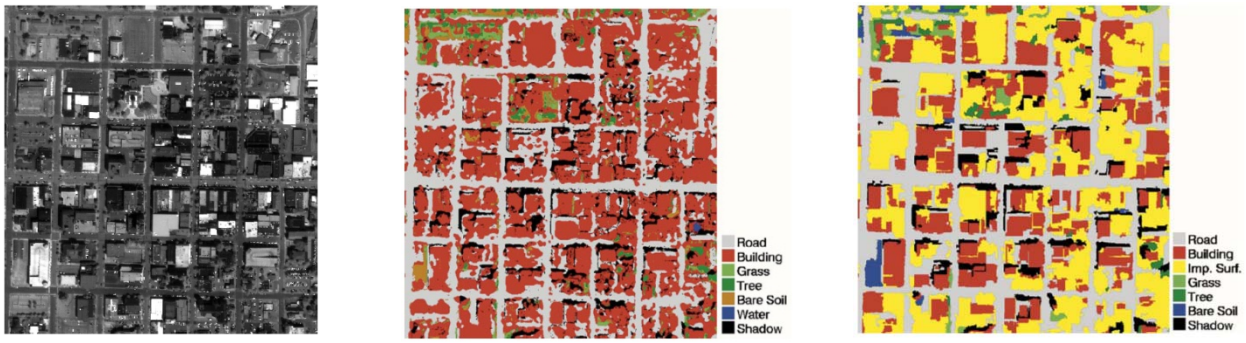


Figure 8: Hierarchical multilevel segmentation showing three different sets of paramaters in (a) to (c) used to classify the pixels in (d).

Shackelford presents a similar method in [4] by assigning several spectral and textural features to each pixel and producing a maximum-likelihood classification map. He then uses an object-based fuzzy classification to identify individual structures. Figure 9 shows this process on an image of downtown Columbia, Missouri.



Original one-meter resolution panchromatic IKONOS image

Crisp output of maximum-likelihood classification

Object-based fuzzy classification

Figure 9: Shackelford's segmentation process beginning with a pixel-based maximum-likelihood classification followed by an object-based fuzzy classification.

Shackelford's method identifies buildings with 76% accuracy in this example. Most automated segmentation methods offer similar classification rates due to the inherent complexity of the problem. Automated techniques still tend to be limited to small test areas and are not applicable on a general scale. These numbers will improve as more research gives new insights, but for now hand segmentation remains the most reliable way to obtain the object classification required for the Text to Sketch application. Figure 10 shows the crisp classification that hand segmentation provides, even when buildings are hard to identify due to surrounding vegetation.

Figure 10: Hand segmented image of Pensacola, Florida.

Data Collection

Testing the Text to Sketch system requires that we generate several descriptive scenarios in areas where we have accurate ground truth data from satellite images. We have segmented Columbia, Missouri and recorded descriptions for several walking paths. Two of the most distinct descriptions involve walking along a road downtown and walking through the University of Missouri campus. The downtown description is far easier to sketch, as it uses phrases that are easy for the system to parse (Figure 12). This is partly because this description targets the Text to Sketch application specifically. It uses very specific phrases such as "I see a moderately small rectangular building close to me that is mostly to my left but partially forward." It also describes buildings in ways that may be difficult for people on the ground, such as "L-shaped." In the campus description, the phrasing is far more natural, although less descriptive in a useful sense (Figure 11). Descriptions may be inaccurate or refer to things that do not show up in segmented imagery such as building height, or decorative architecture. This makes it much more difficult for the Text to Sketch system to build an accurate sketch.

"Ok, I am leaving Engineering Building West. There is a large building in front of me across the street. I am crossing the street at the crosswalk. I am turning to the right, facing south. The building to my left has about three stories, at least one block long. There's a large construction area to my right. The road is completely, um, under construction at this three-way intersection. I can see smokestacks directly to my left. And there is a small building across the street to my right. There is more construction on my left, and to my left I can see a large domed building in the distance. On my right, I am passing a medium sized parking garage. It has roughly three stories. And there is a small parking lot on my left. There is a building on my left somewhat back from the corner, roughly three stories high. I am approaching a four-way intersection. There is a parking garage directly in front of me, about four stories tall. And I am turning left at the intersection. The building on my left is a medium sized building, again roughly three stories tall, and it appears to be rectangular. I am approaching a three-way intersection. On the corner to my front right is a medium sized building approximately four stories tall with a clock and a circle drive. I'm walking toward this building now…"

Figure 6: Campus description.

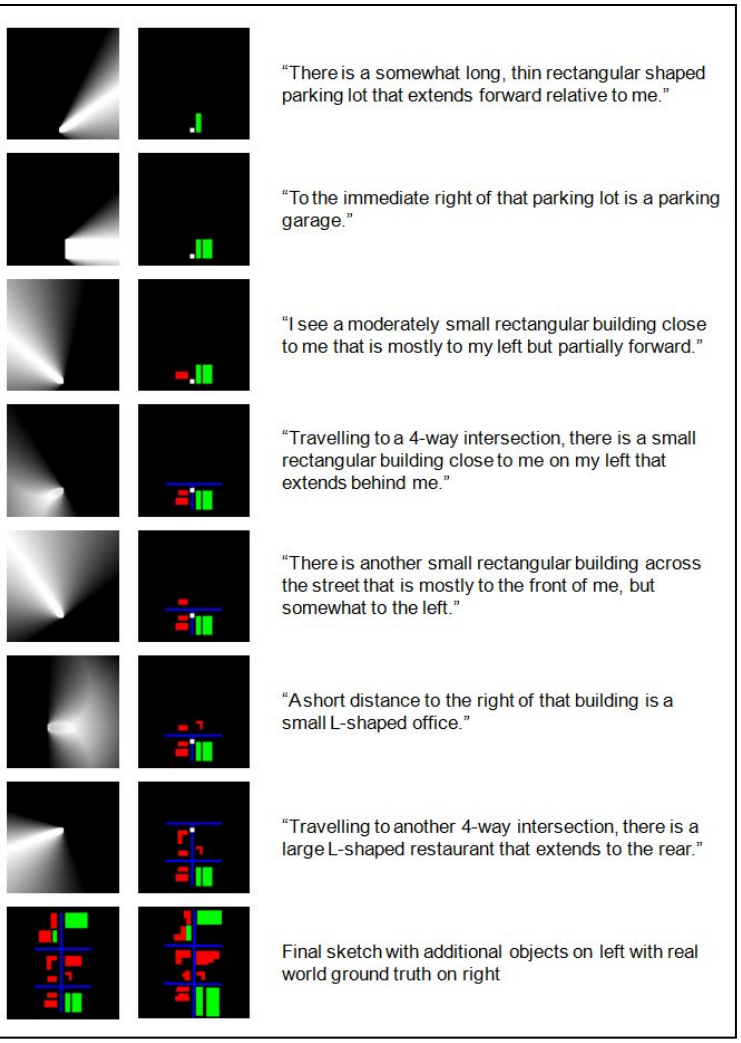


Figure 12: Downtown description.

Comparing these two descriptions shows some of the limitations of the current Text to Sketch system. While the downtown description demonstrates that the system works from a theoretical standpoint, the campus description shows that it is not yet entirely practical. Many of the necessary improvements must come from the language parser to understand descriptions that sound more natural. It could also be useful to assign additional features to buildings in a segmented image. Architectural descriptions such as building material, height, domed roofs, or clock towers are things that a pedestrian can identify more clearly than the overall building size or shape. These features are difficult to discern from a satellite image however, and may require an additional source of information.

Conclusion

The concepts of spatial reasoning and image matching are especially useful in the fields of computer vision and geospatial intelligence. A system that can identify a real world location from a verbal description has many applications, many of which we can only imagine. It can supplement existing global positioning systems by providing more localized information and can provide this service when a GPS is unavailable. However, translating a verbal description into a real world location is no trivial task. It requires that we have a deep understanding of the human mind and the processes it goes through to build a sketch from a verbal description.

Our current understanding of the problem breaks the process into several smaller modules, each with a dedicated field of research. The Text to Sketch module may be the newest and least explored of these modules and shows great promise. The accuracy and scope of its input and reference data is the primary source of error and will likely be the focus of future developments. Automated image segmentation methods continue to improve, but will likely never reach the same level of accuracy as hand segmented images. In addition, the amount of data available in a satellite image may not be enough to make the Text to Sketch system practical to a pedestrian user. A complete system in this sense must include contextual information about an area that a satellite image cannot provide. Despite this shortcoming, the current method of building a sketch is an invaluable stepping-stone. From this, we obtain great insight into the realm of human understanding and spatial reasoning.

References

[1] I. Sledge and J. Keller, "Mapping Natural Language to Imagery: Placing Objects Intelligently," *IEEE Proceedings, International Conference on Fuzzy Systems* (FUZZ-IEEE), 2009 (accepted, in press)

[2] M. Fauvel, J. Chanussot, and J. A. Benediktsson, “Decision Fusion for the Classification of Urban Remote Sensing Images,” *IEEE Transactions on Geoscience and Remote Sensing,* vol. 44, pp. 2828-2838, 2006.

[3] L. Bruzzone and L. Carlin, “A Multilevel Context-Based System for Classification of Very High Spatial Resolution Images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, pp. 2587-2600, 2006

[4] A. K. Shackelford and C. H. Davis, “A hierarchical fuzzy classification approach for high-resolution multispectral data over urban areas,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, pp. 1920–1932, 2003.

This work is funded by the U.S. National Geospatial Intelligence Agency NURI grant HM 1582-08-1-0020.