Learning from Main Streets

A Machine Learning Approach Identifying Neighborhood Commercial Distircts

Jean Oh[1], Jie-Eun Hwang[2], Stephen F. Smith[3], and Kimberle Koile[4]
1,3: School of Computer Science, Carnegie Mellon University

2: Graduate School of Design, Harvard University

4: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

Key words: Main Street Approach, Community Development, Artificial Intelligence, Machine Learning, Active Learning Algorithm

Abstract: In this paper we explore possibilities for using Artificial Intelligence techniques to boost the performance of urban design tools by providing large scale data analysis and inference capability. As a proof of concept experiment we showcase a novel application that learns to identify a certain type of urban setting, Main Streets, based on architectural and socioeconomic features of its vicinity. Our preliminary experimental results show the promising potential for the use of machine learning in the solving of urban planning problems.

1. INTRODUCTION

The recent progress in network technologies and computational power has opened up new opportunities for Artificial Intelligence (A.I.), enabling the use of complex data analysis algorithms on a large amount of data available from online information sources. Machine learning, in particular, has been successfully used in various practical problem domains including text categorization (Joachims 1998; Yang 1999; Sebastiani 2002) and computational biology (Baldi and Brunak 1998), demonstrating competitive performance accuracy against human experts. In this paper we explore the

possibilities of using A.I. techniques in urban design decision support systems.

Levy (1997) summed up the need for urban design[1] in two words: interconnectedness and complexity. Urban design is a complicated decision-making process that involves multiple entities of various interest and constraints. The heterogeneity of urban resources and diversity of involvement raises serious demands for sophisticated communication between government, civic institution, engineers, real estate developers, communities, etc. The communication process, especially with the local community has become more important with the growth of civic rights and public activism (Innes 1998). Moreover, each city has very unique situations and problems. The explicit and insightful examination of the locality is resolved into contemporary major urbanism issues: regional planning, comprehensive planning, and strategic and sustainable development. Therefore, the analysis and identification of existing (and prospective) urban context in a systematic way is the most essential task in the design process.

The Geographic Information System (GIS) is one of the most popular design and decision support tools, facilitating basic data and diverse analysis platforms (Batty, Dodge et al. 2000). Although GIS is a powerful tool it is still limited lack of sophisticated data analysis and inference capability. Our research goal is to add intelligence to design support tools such as GIS, and provide support for designers' decision making processes by effectively analyzing relevant collections of information.

A human expert makes use of a pool of accumulated knowledge from the past when solving a new problem. Let us assume that human experts are *rational* users, i.e., experts who make consistent decisions in order to find the optimal solution that would yield the maximum expected future rewards. Given the assumption of rational users we naively hypothesize that a computer system can learn an expert's decision making knowledge.

Representing such knowledge in a structured form, however, is a difficult problem. One of the earliest approaches was a rule-based expert system. A set of rules of thumbs from domain experts defines intelligence in such systems. In practice, a fundamental limitation of an expert system has been the lack of learning capabilities and the fact that knowledge generally must be programmed into them.

A more general approach is to learn a mapping function from an input to an output. In this paradigm a set of descriptive properties called "features" is presented as an input to a decision making problem, and an output is

[1] In this paper, we use the term "urban design" interchangeably with the term "urban planning." Although urban design and urban planning are distinct disciplines, the roles and professions often are commonly overlapped in practice. Since we address the architectural level of space and built form eventually, we consider this study to be in urban design.

represented as a discrete-valued label, e.g. commonly a binary outcome of "yes" or "no". This paradigm is powerful enough to capture a great many real life decision making problems. During the past decade numerous efficient machine learning algorithms have been invented and applied to problems in a wide variety of domains, e.g., ranging from medical decision support systems to spam email filters.

We take a machine learning approach towards solving urban design problems. We hypothesize that there exists a class of urban design decision making problems that can be formulated as a machine learning problem given a set of specific assumptions. Our first assumption is that a human expert has a certain decision making criteria that is represented in domain specific knowledge. The goal is to train the system to learn the expert's decision criteria. The system's model of domain knowledge, e.g., a set of features, is bounded by data availability. So, we further assume that the system has access to domain specific information sources. The growth of the internet has brought easy access to vast amount of domain specific knowledge. For instance, a building' structural information and its land use were publicly available for our experiment.

Although the field of A.I. has been drawing increasing attention in the urban design domain, little has been done to demonstrate true benefits of A.I. techniques in this problem domain. We showcase our proof of concept experiment by building a software program that can identify a certain type of urban setting, Main Streets. Our preliminary results show promising potential for the use of machine learning in this problem domain.

Considering the interdisciplinary nature of this paper we include introductory sections for readers from diverse backgrounds. Section 2 addresses our target research issues in urban design, and Section 3 describes various machine learning paradigms that are used in our research.

2. URBAN DESIGN PROBLEMS

The integrated perspective of form and function in urban studies is not an innovative notion. In fact, it has been the core subject of urban matters for a long time (Sitte 1965; Needham 1977; Rapoport 1990; Batty and Longley 1994; El-Khoury, Robbins et al. 2003). Previous work, however, has primarily focused on one dominant aspect of either form or function from a particular view point, e.g. architecture, psychology, sociology or economics. Furthermore, the range and definition of form and function varies according to diverse disciplines. For instance, while architects regard form as three dimensional shape of space and building components in the intimate detail, economists rather view it as two dimensional shape of cartographic plane at

the regional or national scale. Architects consider function as activities in individual building spaces and the in-betweens, whereas policy makers consider function as performance of parcel or zone in the whole system of the city.

Resolving multiple views has been an important issue in urban design decision making. The urban design profession contributes to shape the city through designing physical structures; however, it has generally been an execution of form-based policy in this respect (Krieger 2004).

Recognizing the importance of considering interdisciplinary aspects of a problem, urban designers have developed methodological frameworks to investigate urban morphology in a manner that combines interdisciplinary aspects (Bosselmann 1998). Our research contributes to this effort, by applying AI techniques to develop improved representations and methods for reasoning about urban design issues in an integrated fashion. In this paper we focus on an important methodological framework, *typology*, which represents the understanding of urban settings by classification based on present architectural and socioeconimic elements. Specifically, we aim to automate the urban typology process by formulating it as an A.I. classification problem.

2.1 Main Streets

We have conducted our study on an illustrative example of urban design process of *Main Streets*. A Main Street represents a commercial district in the center of a residential area. Throughout the history of urbanization in the United States, Main Street has evolved to serve goods, services, and public activities in a residential neighbourhood. The prosperity of Main Streets, however, declined with the rapid spread of big-box style national chain regional malls along highways during the suburban sprawl era (Isenberg 2004). Along with the widely raised criticism about losing local economic sustainability and cultural identity, revitalization of Main Streets has become the core issue of community development. Many American cities have developed innovative policies intended to restore prosperity and vitality to downtowns. Currently, over 1,000 Main Street communities have adopted "the Main Street approach", which includes a four-point program: organization, promotion, design, and economic restructuring (Robertson 2004).

The Main Street approach is, as many urban design subjects are, usually pursued in partnership between public and private sectors. For instance, the city of Boston has the Main Street program in the department of neighbourhood development in the city hall. Such a team of public sectors collaborates with private sectors, e.g., local Main Street directors who are

usually elected or hired by the community. In a city or regional level, the Main Street is a vital strip within the whole vessel network of the city. At the same time, the Main Street is the center of the local area in a local neighbourhood level. Since each Main Street has unique characteristics and problems identified by the neighbourhood in which it belongs, it is important to understand and identify the local context of the community. Additionally, along with the consideration of historical preservation, the Main Street approach conveys reallocation of existing architectural and socioeconomic resources, as opposed to urban renewal, in the neighborhood.

Accordingly, Main Streets raise an important issue that stems from the complexity of communications among multiple actors. The set of actors involved in Main Street design process includes city officials, local directors, design professionals, communities, developers, investors, etc. The key to a successful Main Street design lies in resolving diverse interests and constraints of multiple actors from architectural, social, economic, and historical perspectives. We propose a systematic way to work out the "multiple views" problem of urban typology by providing an intelligent decision support system that can learn various actors' typology decision criteria.

2.2 Urban Typology

Urban typology (Caniggia and Maffei 1979; Krier and Rowe 1991) has long been a study of subject yet little or no attempts have been made to automate this process. In general, urban typology analysis is a time consuming task that requires complex data analysis and field studies.

For instance, the ARTISTS (Arterial Streets Towards Sustainability) project in Europe was developed to identify a set of types of streets in order to provide better insights to urban planners and economists. This 2.2 billion euros budget project involved 17 European countries and took three years to classify five categories of streets (Svensson 2004). Although their classification rules were drawn from statistical analysis, human experts were the main sources of information for this project. The experimental results show how they classified 48 streets into 5 categories based on their decision rules. Our attempt is to carry out a similar classification task on Main Streets but in an automated way using machine learning techniques.

Moreover, a systematic approach can support a designer's typology view by providing a theoretical justification for her classification decision criteria which is a perspective envision of larger and more complex urban settings based on her unique design concepts. In practice, urban typology studies have been presented in subjective exhibitions without comprehensive denotations of the decision criteria. A result of typology study is often

represented as symbols or specific representative cases. In this paper we propose the use of machine learning in representing such decision making knowledge.

3. MACHINE LEARNING

This section provides a high level overview of machine learning techniques that are used in our research. Machine learning refers to a software system that self-improves through autonomous acquisition and integration of knowledge. In other words, machine learning is a system that builds knowledge from past experience to improve expected future performance (Mitchell 1997). For example, historical medical records form a knowledge source that can be used to diagnose new patients. A doctor makes diagnostic decisions based on a patient's symptoms and a careful analysis of historical medical records. Similarly, machine learning can be used to make predictions for new findings based on what has been observed. They are typically used in two ways -- for classification and for clustering.

A *classification* is a learning task of mapping a set of input variables into a set of outcome variables also referred as "classes". A *classifier* is a system that learns such mapping functions from a set of training examples whose classes are correctly labeled. A classification is a type of *supervised* learning in which the set of output variables are well-defined, e.g., text categorization. It is "supervised" because it requires an expert who can map each training data point to its correct class label. e.g., in order to learn how to classify a vicinity of Main Streets the system needs to observe some set of examples of Main Streets and non Main Streets. Further technical details of a classification are discussed in section 3.1.

A *clustering*, on the other hand, is a type of *unsupervised* learning in which there is no distinction between input variables and output variables. Thus, labeling is not necessary in clustering. For instance, a clustering algorithm can group data into a set of clusters each of which contains only those data points that are similar to one another. In our experiment, we use clustering algorithms to pre-process low level data, e.g., grouping buildings into a set of neighborhoods.

3.1 Classifiers

There are two types of classifiers: generative classifiers and discriminative classifiers (Mitchell 1997). Let X be an input vector of n features, $X = \{X_1, X_2, ..., X_n\}$ where X_i denotes a random variable representing the i^{th} feature. Let Y be an output variable. Let us also denote x_k

be an instance of X, y_l be an instance of Y. Learning of a classifier is approximating an unknown mapping function $f: X \rightarrow Y$, or equivalently estimating a probability of Y given X, denoted by $P(Y|X)$. Generative classifiers assume that input data is generated from a certain statistical model of input variables, i.e., $P(X)$. Based on this assumption, such classifiers estimate $P(X)$ and $P(X|Y)$ from training data and apply Bayes rule to compute $P(Y|X=x_k)$.

Equation 1 Bayes rules

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

$$P(Y = y_i|X = x_k) = \frac{P(X = x_k|Y = y_i)\, P(Y = y_i)}{\sum_j P(X = x_k|Y = y_j)\, P(Y = y_j)}$$

Discriminative classifiers, on the other hand, make no assumption on underlying distribution model of X. Instead, P(Y|X) is directly estimated from training data. In our experiment we mainly used a Support Vector Machine (SVM) classifier. An SVM is a discriminative classifier that learns an optimal classification boundary that minimizes training set error (Burges 1998). The sets of data points near the separation hyperplane are called support vectors. The Figure 1 shows a simple noise-free SVM. An optimization objective of the linear SVM in Figure 1 is maximizing the separation margin width M, which is equivalent to minimizing $w.w$.

In our initial experiment, we tried a set of classifiers to determine the best-fitting classifier in our particular problems. Among a set of Decision Trees, a Naïve Bayes classifier, a k-Nearest Neighbors (kNN) classifier, and an SVM classifier, an SVM classifier best performed (Yang 1999)[2]. In general, SVM is considered one of the best performing classifiers in many practical domains. Despite SVM's high quality performance users outside A.I., such as designers, tend to prefer Decision Trees or generative models due to the fact that their results are more comprehensible. Another limitation of the discriminative approach is handling of outliers, i.e., data points far from training set. Since discriminative classifiers do not attempt to model distribution of input variables it is hard to recognize outliers.

In order to overcome general drawbacks of discriminative approach an outcome from an SVM classifier can be presented in alternative ways. For instance, we train a decision tree that is equivalent to the learned SVM classifier in terms of classification results over the test set. That is, after training an SVM classifier using a set of train data, the system labels the remaining set of data with SVM's prediction. Finally we train a decision tree

[2] Due to limited space we omit formal definitions of various classifiers and refer to Yang's work (1999) that extensively evaluates various types of classifiers.

using the original set of train data plus the remainder of data labeled by the learned SVM. Similarly, we can train a Bayesian classifier to take advantage of benefits of generative models.

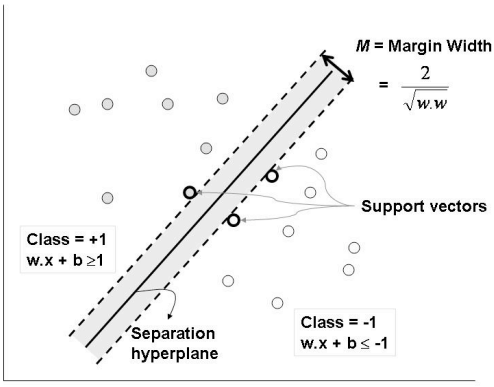

Figure 1. Linear Support Vector Machine

3.2 Active Learning

In some problem domains, cost of labeling is not an issue, e.g., historical medical records. In our problem domain, however, labeling is an expensive procedure. We project a typical typology analysis into a simplified three-step process: data analysis, field study, and decision making. Among these three steps, the field study is the most expensive procedure in terms of both labor cost and time. In order to minimize labeling cost we favor learning algorithms that work well with a relatively small number of training examples.

An *active learning* algorithm enables the system to actively choose a next set of training examples to be labeled. The intuition is to identify the most informative data points by utilizing a larger amount of unlabeled data as a learning guide at no cost. For example, an active learner would select the data points for which it has low classification confidence (Tong and Koller 2000).

A *semi-supervised* method also utilizes distribution of a large amount of inexpensive unlabeled data to guide supervised learning. For example, the co-training method (Blum and Mitchell 1998) learns two classifiers using disjoint sets of features, i.e., two different views over the same set of data, and admits only those predictions upon which both classifiers agree. A more recent approach includes incorporating clustering into active learning (Nguyen and Smeulders 2004). Using prior data distribution their system first clusters data and suggests cluster representatives to the active learner.

Their algorithm selects not only the data points close to classification boundary but also representatives of unlabeled data.

4. MODELING

Modeling an urban typology as a machine learning problem is based on two important assumptions: 1) a set of relevant features that define an input to a learning algorithm are known in advance, and 2) data that describe the features are a well-structured set of vectors. Applying machine learning algorithms to a well defined set of data is a straightforward task. However, a major difficulty of formulating urban typology into a machine learning problem resides in feature space modeling and compiling a set of relevant data.

The human experts' elicitation of relevant features is often vague and incomplete. We exemplify a modeling of feature space in Figure 2. This example depicts the feature dependency graph that represents a perception of *publicness*. Publicness is a meaningful concept in urban design and relates to how people perceive whether a given urban component is public or private. We modeled this example based on a survey that was conducted on both domain experts and non-experts. Although this example does not directly address the problem of Main Streets the features in the graph, such as Massing, are commonly used as urban decision criteria, and thus they are relevant to our discussion.

Among these features the entries that are drawn in boldface in Figure 2 are the set of features that users considered important in decision making. Because the system can only recognize well-structured data, e.g., features stored in databases, only the features shown in grey are included in our model. This example illustrates our modeling assumption that domain experts' model of relevant features are often abstract semantic concepts that *depend* on descriptive features that are available in low level databases.

Massing, for instance, is a feature that differentiates buildings by their structural size information. In our information sources Massing is represented as multiple features, height, area, periphery, distance to nearest neighbor, etc. Our survey result also reveals the existence of hidden features that are completely isolated from what is available in low level database. These hidden features were denoted by *intangible* features in the picture, e.g., features related to "Use Patterns".

We learn from this example that a majority of features in a human user's model are abstract concepts, whereas the system only has access to low level databases. We make a specific assumption that abstract concepts that human experts consider relevant in fact depend on low level features in databases.

We also assume that the system has access to such domain specific information sources. The challenge then is to infer the mapping from low level features to abstract concepts.

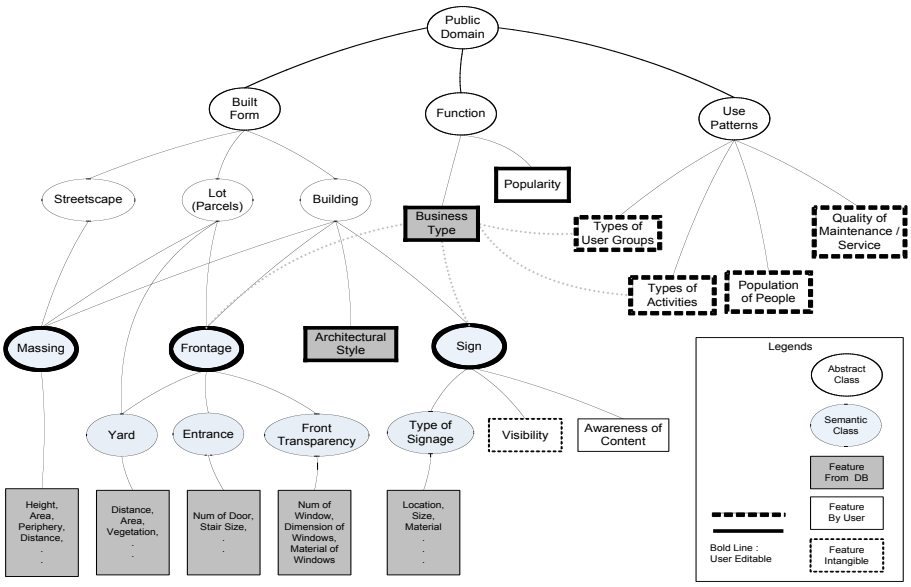

Figure 2 Features for determining publicness of a urban component

5. EXPERIMENT

This section describes a preliminary experiment carried out to verify our hypothesis. We chose the specific problem of identifying a certain type of urban setting, Main Streets, based on architectural and socioeconomic features of its vicinity. The criteria of classifying a commercial district varies from city to city, thus it is hard to find a generalized set of rules to distinguish Main Streets from rest of districts. Therefore, we aim to establish the customized decision criteria of identifying Main Streets over the proposed data model for a particular city.

Most machine learning algorithms expect data to be a well defined set of tuples, but in reality this is rarely the case. For example, if data is stored in a relational database with multiple tables the data must be merged into a giant single table. Building an inference network from a relational database is an interesting area of research (Getoor 2001) and we also anticipate that our future work may be in this area. For the sake of simplicity we assumed that we already have the data formatted into a set of tuples in this experiment.

5.1 Data Pre-processing

We investigated Main Streets in the city of Boston for this study (Figure 3). Boston provides an ideal testbed for evaluation because a complete set of districts were already identified as Main Streets by field experts. We used relational database tables exported from GIS information sources that are available from the city of Boston. The data was then pre-processed to get it into a format suitable for general classifiers.

Initially we started with two database tables: buildings and parcels. Note that a data entry in these tables represents a building and a parcel, respectively, whereas our target concept, Main Streets, is defined as a district which is usually composed of hundreds of buildings and parcels.

First, we applied unsupervised learning methods to group buildings and parcels into a set of candidate districts. We used a single-linkage clustering algorithm in which every data point starts with a separate cluster and merges with the closest neighboring cluster until a given proximity threshold is satisfied. The proximity threshold was chosen empirically to generate reasonable size clusters.

Our algorithm for identifying district candidates consists of two clustering steps. Since the backbone of Main Streets is a strip of commercial buildings we first clustered buildings that are associated with commercial land use code in order to retrieve strips of commercial buildings. For this research study small clusters that contained less than 5 commercial buildings were filtered out. In the second step, the commercial strips identified in the first step were treated as a single cluster when the second round of clustering started, i.e., the set of initial clusters in the second round was the union of commercial strips, non-commercial buildings, and all of parcels. The number of buildings and parcels in the resulting district candidates were in the range of hundreds.

For simplicity, we used Euclidean distance between the two centers of buildings as the distance measure. In order to refine cluster boundaries we need to incorporate more accurate separator data, e.g., geographic obstacles such as mountains or rivers, and man-made obstacles such as bridges and highways. This will be an interesting topic for a future work.

Using a raw data set containing 90,649 buildings and 99,897 parcels (total around 190,000 data points) our algorithm identified 76 candidate districts. Each candidate cluster corresponded to one data row for a classifier, and aggregated characteristics of a candidate cluster, such as average height of the buildings, were used as features.

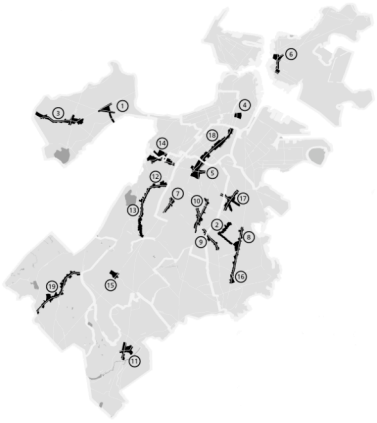

Figure 3 Main Streets in Boston, Massachusetts

5.2 Finding Main Streets

In the data pre-processing step we form a set of candidate districts of Main Streets using unsupervised learning algorithms. We then use a supervised learning method to find Main Streets among this set of candidates. Table 1 lists the set of features we used to train a classifier[3].

Table 1 Main Streets features

Aggregated features	Number of buildings
Ratio	land use (commercial/residential), parcel business type
Average	building height, perimeter, lot size, stories, built year, renovation year, shape length, shape area, gross area, living area

Labeling is an expensive process in this domain because labeling one district requires thoughtful analysis of huge amounts of data and it involves field study. This cost-bounded domain constraint leads us to favor learning algorithms that work well with relatively small number of training examples.

An active learning algorithm reduces the number of training examples by actively choosing the next training example to be labeled. We selected Tong and Koller's approach over SVM (Tong and Koller 2000). The basic idea is to suggest data points that are near the separation boundary, which is quite intuitive and has also proven to be very effective in other practical domains such as text classification. We extended the SVM-KM Toolbox (Canu et al. 2003) to implement the active learning strategy.

We also tried incorporating pre-clustering (Nguyen and Smeulders 2004) to find the initial samples to be labeled. This technique, however, didn't have

[3] We used all numeric features in the buildings and parcels database, some of which are not included in Table 1.

a significant impact on performance in our experiment mainly because the size of unlabeled data was not large enough (After pre-processing we had only 76 district candidates). We would expect higher impact on performance if we had a larger set of data.

5.3 Experimental Results

We have built an intelligent urban design decision support system that aims to solve a complicated decision making problem of urban typology, Main Street. First, we evaluate a general classification performance of our system to prove that the typology of Main Street can be efficiently developed by a machine learning approach.

We used precision, recall, and their harmonic mean as evaluation measures. In our example, precision *p* is the ratio of the number of correctly predicted Main Streets to the total number of positive predictions. On the other hand, recall *r* is the ratio of the number of correctly identified Main Streets to the total number of Main Streets in Boston. Because the two measures are in inverse relation their harmonic mean is often used as a compromising measure. F1 measure, which is a harmonic mean of precision *p* and recall *r,* is defined below.

Equation 2 F1 measure

$$F1 = \frac{2pr}{p+r}$$

Since we had relatively small sized data points after pre-processing we used Leave-One-Out-Cross-Validation (LOOCV) to evaluate the general performance of Main Streets classifier. LOOCV is a cross validation technique where one data point is left for testing while a classifier is trained using the rest of data points. The LOOCV results in Table 2 shows promisingly good performance by achieving high F1 measure of 0.8. The results read that the system made 6 correct predictions out of every 7 trials, identifying 76% of Main Streets.

Table 2 Leave-One-Out-Cross-Validation Result

	Precision	Recall	F1 measure
LOOCV	0.842	0.762	0.800

We also compared the performance of the active learning strategy to the performance of the random learning strategy. Under the random learning strategy the system also learns an SVM classifier by incrementally taking more training examples. Whereas the active learning strategy takes advantage of distribution of unlabeled data in selecting a next data point, the random learning strategy chooses an arbitrary data point. We evaluated the performance of the two approaches in terms of their learning speed.

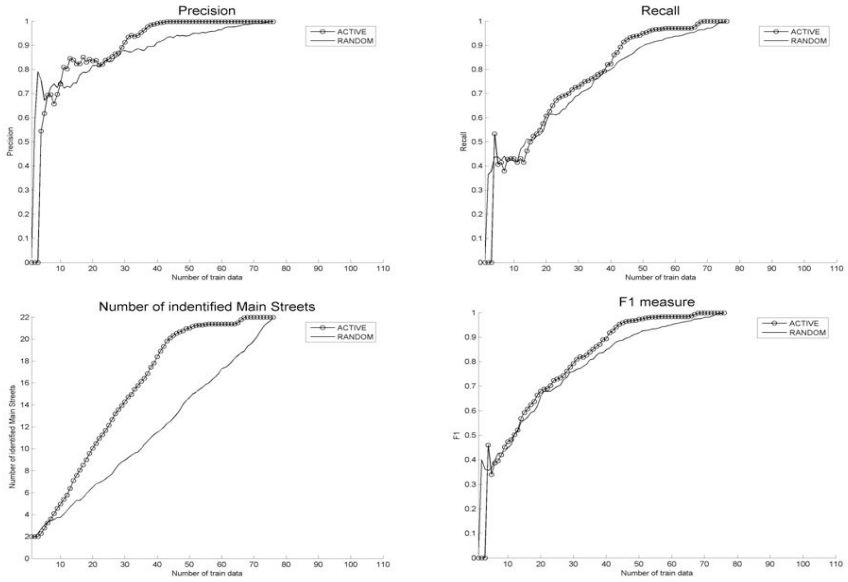

Figure 4 Active learning strategy vs. Random learning strategy

Figure 4 shows the performance of active learning strategy and random learning strategy. We included the number of identified Main Streets (the lower left in Figure 4) which also clearly demonstrates superior performance of the active learning algorithm. The experimental results in Figure 4 are average performance of a set of 20 independent trials.

As shown in Figure 4, the results illustrate the benefit of learning, i.e., a steady progress in performance proportional to the amount of experience, such as a number of training examples (X axis). First, we notice that the precision of our system under both active and random strategies reach nearly .7 which is a reasonably high precision level after observing approximately 10 training examples. The level of recall also reaches .5, i.e., a half of Main Streets have already been identified, given 20 training examples. The experimental results first indicate that finding Main Streets is a class of urban design decision making problems that can be solved by a machine learning approach. The results also show that the active learning algorithm significantly[4] outperforms the random learning algorithm, achieving high classification accuracy after given a relatively small number of examples.

Our initial hypothesis was that there exists a class of urban typology problems that can be modeled as a classification problem. Our current study shows that classifiers perform very well on identifying Main Streets even with an incomplete set of features. Especially using an active learning

[4] This is statistically significant with a strong evidence of p-value = 0.01.

algorithm our system can cleverly choose better samples to be labeled, outperforming a random selection model significantly.

6. CONCLUSION AND FUTURE WORK

Urban design is a complex decision making process that must compensate for multiple stakeholders' various interests with respect to physical, social, and economic constraints. Urban design professionals need to collect and thoroughly analyze large amounts of data in order to make robust plans towards long-term goals and to communicate with diverse stakeholders. This is normally a careful and time-consuming task, due in part to limited financial resources but also because design decisions often generate cascading effects contingent on both pre-existing physical urban structures and future design decisions.

Although there is growing interest in using A.I. in urban design and planning this community remains a field dominated by human experts. Recent catastrophic disasters such as hurricane Katrina, however, have underscored the need for increased automation and more efficient urban design processes. For example, finding good locations for temporary housing is one of the most urgent decision making tasks in post-disaster urban planning that would benefit from A.I. approach.

In this paper we described our efforts to apply A.I. techniques to urban design problems. As a proof of concept experiment we showcased an application of machine learning that actively learns to identify a certain type of urban settings, Main Streets. Our preliminary experimental results show promising potential for utilizing A.I. techniques in this problem domain. Our current research focuses on more constrained decision making problems particularly in post-disaster settings.

7. ACKNOWLEDGEMENTS

The authors thank Yiming Yang for fruitful discussions. This research was sponsored in part by the Department of Defense Advanced Research Projects Agency (DARPA) under contract #NBCHD030010.

8. REFERENCES

Baldi, P and Brunak S., 1998, *Bioinformatics: The Machine Learning Approach*, MIT Press.

Batty, M., M. Dodge, et al., 2000, New Technologies For Urban Designers: The VENUE Project. Center for Advanced Spatial Analysis working Paper Series.

Batty, M. and P. Longley, 1994, Fractal cities: a geometry of form and function. London; San Diego, Academic Press.

Berges, J.C., 1998, "A Tutorial on Support Vector Machines for Pattern Recognition", *Data Mining and Knowledge Discovery,*2(2), p121-167.

Blum, A. and Mitchell, T, 1998, "Combining labeled and unlabeled data with co-training", *Proceedings of the Workshop on Computational Learning Theory,* p92-100.

Bosselmann, P., 1998, Representation of places: reality and realism in city design. Berkeley, Calif., University of California Press.

Caniggia, G. and G. Maffei., 1979, Architectural Composition and Building Typology: Integrating Basic Building. Firenze, Italy, Alinea Editrice.

Canu, S.; Grandvalet, Y.; and Rakotomamonjy, A. 2003. Svm and kernel methods matlab toolbox. Perception Systmes et Information, INSA de Rouen, Rouen, France.

El-Khoury, R., E. Robbins, et al., 2003, Shaping the city: studies in history, theory and urban design. New York, NY 10001, Routledge.

Getoor L., 2001, *Learning Statistical Models from RelationalData*. PhD thesis, Stanford University.

Innes, E. J., 1998, "Information in Communicative Planning." Journal of the American Planning Association 64(1): 52-63.

Isenberg, A., 2004, Downtown America: a history of the place and the people who made it. Chicago, University of Chicago Press: xviii, 441 p., [2] p. of plates (col).

Joachims, T., 1998, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", *Lecture Notes In Computer Science*, 1398, Springer-Verlag

Krieger, A., 2004, Territories of Urban Design, Harvard Design School.

Krier, R. and C. Rowe, 1991, Urban space. London, Academy Editions.

Lopilato, L., 2003, *Main Street: some lessons in revitalization*, University Press of America, Inc., New York.

Mitchell, T. M., 1997, *Machine Learning*, McGraw Hill, New York.

Needham, B.,1977, How cities work: an introduction Oxford; N.Y., Pergamon Press.

Nguyen, H. T., Smeulders A., , 2004, "Active learning using pre-clustering", *Proceedings of InternationalConference on Machine Learning*.

Rapoport, A., 1990, The meaning of the built environment: a nonverbal communication approach. Tucson, University of Arizona Press.

Robertson, K. A., 2004, "The Main Street Approach to Downtown Development: An Examination of the Four-point Program." Journal of Architectural and Planning Research 21(1): 55-72.

Sitte, C., 1965, City planning according to artistic principles. New York, Random House.

Sebastiani, F. 2002, "Machine learning in automated text categorization" *ACM Computing Surveys*, 34(1), p. 1-47.

Svensson A., 2004, "Arterial Streets For People", Technical report, Lund University, Department of Technology and Society, Sweden.

Tong S. and Koller D., 2000, "Support Vector Machine Active Learning with Applications to Text Classification", *Proceedings of 17th ICML*, p. 999-1006.

Yang, Y. 1999, "An Evaluation of Statistical Approaches to Text Categorization", *Information Retrieval*, Springer, 1(1-2) , p. 69-90