Learning Qualitative Spatial Relations for Object Classification

Tristram Southey and James J. Little

Abstract— **In this paper, we describe research into learning a model of the general 3D spatial relationships between objects found in human environments. Our approach trains a maximum entropy model using the qualitative spatial relationships between all the objects in several environments and produces a model of the underlying spatial regularities between the objects. It is tested as a purely spatial object classifier on the task of recognizing hundreds of objects based on their position relative to the other objects in their environment. We also introduce a novel source of data about object arrangement in human environments from the commercially released computer game Elder Scrolls 4: Oblivion.**

I. Introduction

A better understanding of the structure of human environments and the ability to make predictions about object type based on that structure would be useful for robot problems such as object and scene recognition and object search and retrieval. Our work focuses on the problem of learning a general spatial model of houses from a large synthetic collection of 3D data describing object types, locations and shapes found in many houses. Our goal is to create a *maximum entropy model* of the qualitative spatial relationships between objects and demonstrate its applicability for object recognition based on scene contextual cues.

With this model we train an object classifier based solely on the relative spatial positions of objects in an environment. Our objectives are to demonstrate that spatial data can be used for object classification, that probabilistic models are a good foundation for spatial classification and that qualitative spatial relationships can provide a basis for learning the spatial structure of human environments. Our model is intended for other task, such as identifying expected object position or creating qualitative descriptions of an object's position, though in this paper we will focus on the simpler sub-problem of object spatial classification.

Much of this work is founded upon ideas from *qualitative spatial reasoning* (QSR). A primary goal of QSR has been to find a way to provide computers with an approach with which they can reason about human spaces, and so-called common sense human spatial concepts. QSR aims to take the quantitative descriptions of an environment that would be traditionally available to a computer or robot and allow it to better understand how a human would perceive that environment's spatial characteristics.

One serious issue that has limited research in spatial object classification has been the lack of any appropriately large or accurate source of data on the position, type and shape of all the objects found in houses. Acquiring such data at present would be extremely time consuming, expensive and invasive. However, we have discovered an interesting and adequate source of synthetic data from a commercial video game called Elder Scrolls 4 [3]. The game contains hundreds of human habitations in many different styles and modeled in great detail by the game designers. Through reverse engineering we were able to acquire this data and use it for testing our work.

Our goal is to create a foundation for future work on this topic when natural data is more readily available. With the increasing use of RFID (radio frequency identification) tags in commercial products, the growth of user labeled object databases such as LabelMe [24] and improvements in 3D shape measuring devices like stereoscopic cameras and lidar, we will someday have access to large quantities of accurate, or semi-accurate, quantitative data about the spatial object structure and makeup of human environments.

A. Qualitative Spatial Learning

A key difference between the work presented in this paper and that of the majority of qualitative spatial reasoning research is that it is not concerned with reasoning about the spatial relationships of objects but rather with learning using the tools of QSR. Reasoning is a process of taking data and applying an understood set of axioms to arrive at some result. With learning, the axioms are either unknown or only partially known initially and must be discovered through some other process. In our work qualitative spatial learning or QSL provides a computer with the foundation of QSR, the qualitative spatial relationships, and then allows the computer to determine its own mechanisms by which to reason about the environments described. If the relative position of objects in human environments is determined primarily by their qualitative, not quantitative, spatial relationships, then it makes sense that learning a model of how humans construct their environments would be best achieved through the use of qualitative spatial relations. It is important to keep in mind that it is not simply the reduction in complexity from using a discrete, symbolic description of space that makes qualitative spatial descriptions useful; these descriptions can also capture the relevant spatial distinctions used by humans [5].

To better understand what is meant by qualitative spatial classification, let us examine an anecdotal example of how a human might apply it to a problem. A person is sitting on a chair at a table and in front of them are three objects in a row. The object to the right is a knife and the object to the left is a fork. However, the object in the middle is covered and unknown. Given the scene, what type of object is in the middle? If the scene is natural (i.e., not contrived), most

people in this situation would be able to easily determine that it is most likely a plate [23].

This simple example would be a difficult classification for many visual classifiers because plates, while they share a similar geometry, can have a bland or varying appearance depending on their coloration. However, identification of this unknown object as a plate is possible using only relative qualitative object positions because humans have learned a model of the spatial relationships that exist between objects in human environments and how to apply those relationships to the problem of object classification.

Unlike visual classification, we hypothesize that spatial classifiers will frequently be unable to predict an object's type with a high degree of accuracy using only positional information since even humans cannot do it well. In the above example there are thousands of objects that might be in that position. However, spatial classification could allows us to rank our estimates of object type and to use them as a prior to improve some other classifier, such as visual object classification.

II. Related Work

Visual object classifiers have come a long way in the last 15 years and are now capable of accurately recognizing a limited selection of complex objects with high accuracy [1][7]. However, many of the approaches to visual classification are restricted in their ability to classify large numbers of objects and fail on seemingly simple ones when the objects lack sufficiently distinctive appearance data (e.g., a plain white dinner plate). Humans use the structure of scenes to provide their visual classifier with priors on which objects to expect and look for [2]. It has been shown that improvements in the accuracy and scope of visual object classifiers can be achieved by using the spatial layout of known objects in a scene to provide a probabilistic prior on the type of unknown objects [25].

Significant early work in this area was Rimey and Brown's on selective perception for scene classification [23]. Their approach used a Bayesian network model of a tabletop environment to determine where a camera should focus to classify the scene efficiently. The scenes used were very simple, consisting of tables set for a tea party with a small range of object types. Experts supplied the 2D spatial relationships between the objects, an approach that clearly would not be feasible for large-scale models containing hundreds or thousands of objects. More recent work by Neumann and Möller[17] has examined the role of spatial constraints, taxonomy, and compositional relationships in interpreting objects and actions in scenes.

Some visual classifiers improve the accuracy of their classifications by modeling the consistencies in the 2D relationships between objects in the image plane [15][25] or the expected positions of objects within images [14]. The former of these two approaches are the most similar to our own since they are based on relative object spatial relationships. Other approaches ignore the spatial arrangement of objects and improve scene recognition by recognizing relevant objects (e.g., computers in an office or cars on the street) [16]. Part-based classification is an object recognition technique that identifies objects through the consistent spatial arrangement of their parts. This is often necessary for recognizing complex object classes that can vary in appearance but which contain elements in a consistent spatial layout (e.g., cars and faces). Part-based techniques usually learn a model of relative image position of identifiable regions that make up the object's appearance [26][9][1].

The spatial properties of an image can also be exploited for semantic scene classification, which aims to classify what the overall image is showing (e.g., street scene, beach, indoors) and what is the content of the regions of the image (e.g., water, road, sky, ceiling). Most of these approaches use image segmentation to divide the image into regions and then learn classifiers to identify what each region contains. Results can be improved by learning a probabilistic model that describes the likelihood of two regions being spatially contiguous, such as water being close to sand and sky. Some approaches have gone further and based their probability model on 2D qualitative spatial relationships like *above*, *below* and *surrounding* [4].

There are several aspects to our work that differentiate it from these other approaches. We are researching how to create models purely based on the spatial consistencies between objects in houses, without the assistance of visual classifiers. Very little of the vision-based object classification research has used 3D information for describing the spatial properties of objects. Instead such techniques are typically restricted to using 2D spatial distances [25] and directions in the image plane or, at best, attempting to reconstruct the 3D object configuration from the image data [21]. Another difference is that we use qualitative spatial relationships as the basis for classification, which hopefully allows us to learn models based on the same underlying principles on which the environments were constructed. Lastly, our work is trying to classify a much larger variety of objects than most other classifiers. Instead of restricting our investigation to a restricted set of objects whose semantics are well understood, we have opted instead for the much harder task of classifying everything in the environment.

III. Qualitative Spatial Classification

In this section, we describe our approach to modeling the spatial distribution of objects in human environments for classification. As was previously stated, we are using synthetic quantitative data on the spatial layout of objects in many different houses. Each house is treated as its own separate 3D space and contains a set of objects $x_1, \ldots, x_m$, each of which has a type name (e.g., chair, flower pot, or fork), a 3D position, a 3D orientation and a shape. Using this data, our goal is to classify an unknown object in an environment given its position and shape relative to all other objects in the same house given their shape, position, orientation and type. These allow us to find qualitative spatial relationships between the objects (e.g., X is above and touching Y).

We then use these relationships to formulate a probabilistic classification problem using a maximum entropy framework.

Classification requires that we know the unknown object's 3D position and shape in order to identify its spatial relationships with the other objects in the house. Our approach assumes that each object has a discernible intrinsic frame of reference in order to determine relative orientations between them. There are ways of finding an intrinsic frame of reference given an object's shape, such as looking for the longest and second longest interior spanning distances. These ways of finding an intrinsic frame can be used as a basis for determining object's orientation within the scene and the relative orientation of a pair of objects (e.g., X is left of and above Y), though it is not guaranteed to find an intrinsic orientation. Gravity can also supply a vector on which to base comparisons of two objects' relative height. We do not use either of these approaches since our data source automatically supplies the orientation (as yaw, pitch and roll) of each object relative to vectors built into its model. This is not realistic and will need to be dealt with in future work.

Triangle meshes are a standard format for 3D shape in computer graphics applications and are used both in Oblivion and our classification approach. A triangle mesh consists of a series of 3D points that describe the vertices of a polyhedron made up of triangles. The shape of the objects is necessary only for finding the relative distances between the exteriors of objects. Alternatively, the object's centroid point could be used as the basis for distance comparisons but it is a less accurate measure of separation and cannot identify whether two objects are touching each other. We believe that physical contact is an especially useful spatial feature for object classification since object's generally only touch a small number of other objects.

A. Maximum Entropy Modeling

Statistical models are designed to take an incomplete set of knowledge about some process and accurately and succinctly approximate it. *Maximum entropy models* (maxent) are a class of model that has been applied most commonly to natural language processing [19]. Most often, maximum entropy models are from the exponential family. In a maxent framework, the training data is a sample from a distribution over some feature space. The goal of maxent is to approximate the training distribution with the distribution of maximum entropy, i.e., the distribution that is maximally non-committal with regard to missing information [12].

More formally, consider a random process which produces an outcome $y \in Y$ given some context $x \in X$. In our case, x is a spatial configuration of objects in the environment surrounding a target object y. Let $f_1, \ldots, f_n$ be a set of features where $f_z : X \rightarrow \{0, 1\}$. The goal is to approximate π, a probability distribution over X, given $\tilde{\pi}$, an observed empirical distribution of X, such that:

$$\pi[f] = \tilde{\pi}[f] \tag{1}$$

for all features f_z and where $x[y]$ is the expectation of x under the distribution y:

$$\pi[f] = \sum_{x \in X} p(x) f_z(x) \tag{2}$$

$$\tilde{\pi}[f] = \sum_{x \in X} \tilde{p}(x) f_z(x) \tag{3}$$

A model P then is consistent with our observations if and only if it satisfies (1). In a maxent framework, π should also be the distribution of maximum entropy. The entropy of a distribution p on X is:

$$H(p) = -\sum_{x \in X} p(x) \ln p(x) \tag{4}$$

Therefore we want to find π such that:

$$\pi = \arg\max_{p \in P} H(p) \tag{5}$$

The maximum entropy model can be represented as:

$$\pi = \sigma \prod_{z=1}^{n} \alpha_z^{f_z(x)}, 0 < \alpha_z < \infty \tag{6}$$

where σ is a normalization constant and the α_z's are model parameters that each correspond to the weight of a parameter within that model [22].

There are several advantages to using maximum entropy models. Using features they can incorporate interactions between different types of variables. They can provide a probability for each known type of object, rather than only the type of the most likely object, which allows for comparison between object likelihoods and ranking. There are efficient deterministic algorithms that are guaranteed to converge to a maximum entropy probability distribution. These include generalized iterative scaling [6] or improved iterative scaling [19] that iteratively adapt the probabilistic weight applied to each feature for a given outcome. The training times of these approaches grown linearly in the number of features, which allows us to experiment with a wider range of features and examine how different feature sets interact with each other.

The effectiveness of maxent on spatial classification problems is evidenced by recent work on modeling the spatial distribution of bird species in North America [18], which employed maximum entropy models based on geographic and climate data. The results were found to be consistently as good or better than the state of the art.

B. Spatial Features

Selecting the right feature set is important to the success of any classifier. As was mentioned previously, we chose features based on the relative qualitative spatial positions of the objects because we believe they are the basis on which humans originally constructed the scene and can reduce the complexity of the data while still maintaining the relevant information [5].

The basic structure of a relative qualitative spatial relationship between two objects can be expressed as XRY, where X and Y are the type of the two objects being compared

and R is the spatial relationship (e.g., cup leftOf wine). At present, we restrict ourselves to binary comparisons, which reduces the number of qualitative features that are used in the model and simplifies training.

Proximity and *direction* are two properties that commonly form the basis for qualitative spatial relationships. Qualitative spatial relationships can be divided into three groups: distance-based (near, far), direction-based (left of, right of, in front, behind, above, below) and containment-based (inside, outside, surrounding) [10]. Due to limitations in our data and in the algorithms we use to determine qualitative spatial relationships, we only use the distance and direction-based features. However, we have expanded on the range of distance-based features in the hope of capturing additional spatial properties that exist between objects in human houses.

The distance or proximity-based features set contains the tags: *touching*, *near* (arms reach), *mid* (scene level region), and *far*(more than a scene apart). We determine all proximity features by finding the minimum distance $D(X, Y)$ between the exteriors of the triangle meshes describing the shape of object X and of object Y [11]. Each feature then corresponds to a range of values along this distance. The justification and range for each region are given below:

Touching:
Definition: $D(X, Y) <= 0cm$
Justification: Most often contact between objects is the result of gravity and the fact that every object must be in physical contact with at least one surface (barring extraordinary situations). Touch features are useful for identifying the strong relationship that obtains between objects that physically support each other such as food on plates, books on bookshelves and furniture on floors. Due to a limitation in the Oblivion game engine, there are no objects to be found in any containers that completely surround their contents (like drawers or boxes). This is because Oblivion models these containers as their own non-3D space to simplify physics and the games interface. There are, however, many containers that only partially surround their contents, like bookshelves, that are correctly handled.

Near:
Definition: $D(X, Y) > 0$ & $D(X, Y) < 50cm$
Justification: This proximity range is intended to identify objects that are within immediate arms reach of each other since they are more likely to share a common action or activity (such as eating, cooking, writing, etc) than those farther apart and therefore consistently found in groups. This range is particularly useful for grouping together objects at the tabletop level.

Mid:
Definition: $D(X, Y) > 50cm$ & $D(X, Y) < 200cm$
Justification: This proximity range captures objects that belong to a common scene, such as sharing a room or a semantic sub-division of a room and therefore potentially also sharing a common purpose or purposes. This range is useful for grouping together larger objects, like chairs and tables, at the room level.

Far:
Definition: $D(X, Y) > 200cm$
Justification: This final proximity range relates objects that share the same overall environment (i.e., the same house) but are most likely not functionally associated except at the broadest level. This feature is useful for extracting data on the overall object population of a house. For example, it can be used to recognize that some objects in a building would only be found once anywhere (like a stove or a dining table).

The directional feature set we employ uses the tags *above* and *below*, *inFront* and *behind*, and *leftOf* and *rightOf*. Directional features are asymmetrical, (i.e.,$D(X, Y) \neq D(Y, X)$) unlike distance features. Each directional feature, $d(X, Y)$, is defined using the position of the target object X relative to the position and orientation of the compared object Y. The feature consists of one of each of these three pairs of features above (e.g., X right/below/behind Y). This divides space around the compared object Y into 8 separate regions into which object X can fall. The directional comparison is made between the centroid point of both X and Y and takes into account the orientation of Y.

IV. Object Spatial Data

Acquiring a large amount of data about the shape, position and orientation of all the objects in many houses is not feasible at present. Doing so manually would be impractical given the amount of time, money and effort it would require and we do not yet possess the tools to automate the process. However, to test work, we identified a novel source of high quality synthetic data from a recently released commercial video game called Elder Scrolls 4: Oblivion [3]. Figs. 1 and 2 show examples of scenes from Oblivion, a tavern and a dining room. The spatial relationships demonstrated in these scenes show a surprising level of variety and detail. For example, the wine rack in the back of the tavern contains many different types of wine placed into its lattice structure. The dining room table is set with over 15 types of food, with different arrangements of food on the dinner and serving plates. Other complex relationships not shown include tools standing upright in barrels, bowls piled with fruits, and books ordered into sets on bookshelves.

A number of factors make the data from Oblivion appropriate for our work. First of all, there is plenty of material. Oblivion contains several hundred houses, castles, forts, churches, lighthouses, mines and other types of habitations, though in our work we restrict ourselves to places containing the word "house" in their description. Houses appear in

Fig. 1. A tavern from Oblivion. The wine rack in the rear of the bar holds about 30 bottles of wine in 5 different varieties. These were collapsed into a single object type "wineBottle" for classification purposes. The cat-like character behind the bar is the owner.

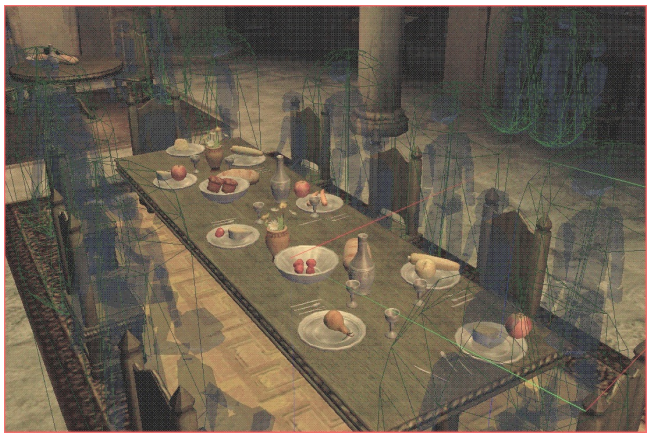

Fig. 2. A large and ornate dining room from Oblivion. In the center of the image is a dining table containing with food and wine set for 10 people. Surrounding the table are chairs, though these are partially obscured by some shapes that show how character models would transition from standing to sitting on the chair.

different architectural and ethnic styles ranging from city to city and none of the houses are simply identical copies of each other. The contents of each house are modeled down to an impressively fine level of detail. Tables in the houses are set with meals, shelves are full of books and ornamentation, bedrooms are laid out with beds, dressers and clothes. Altogether, there are over 1000 different objects modeled for the game. The setting is a medieval fantasy world so the contents of the houses are somewhat antiquated but the complexity of their environments is substantially beyond what is normal in any other game. Fortunately, the designers of Oblivion were good enough to design their game to be easy to modify and customize. The data files can be parsed to extract the layout of the entire game world and there are tools for extracting the object models.

Oblivion is unusual since almost all of the objects contained have no purpose but to make the environments seem realistic. In fact, many of the houses we are using would never even be seen from the inside by the player, existing primarily as an interesting backdrop. This is a very important factor in using video game data since in most games the realism of an area is reduced to meet the artificial requirements of gameplay. For example, in most games, if the player were just supposed to move from one end of a building to the other quickly, most of the rooms in the building would not be modeled. Even if a room is modeled, most of the contents will be *game artifacts*, objects that are primarily significant to the gameplay, like weapons or health aids that would be artificially placed and not appear in the real world. There would only be a small number of background objects to make the environments realistic.

There are some limitations of the Oblivion data set. Due to a feature of the games design, none of the completely enclosing containers (e.g., boxes, chests, or drawers) in Oblivion actually contain objects but bookshelves, not being enclosed, do. In addition to this, at present we are not able to acquire data about the shape of the actual building or its components (e.g., walls, floors, doors, etc) and they are ignored. We hope to deal with both of these limitations in future work. Lastly, despite the high level of detail, there are fewer objects in an Oblivion house than in a real one. However, since this is still early work on this problem, it is appropriate to use a simplified data set. Despite the medieval setting, the objects still have a spatial structure that can be modeled and design of the houses is relatively modern, with many having multiple floors, large rooms and windows and areas like bedrooms, kitchens, and offices.

V. Experiments

The data we acquired from Oblivion for testing and training comes from 197 houses that were selected randomly containing a total of 11659 objects with 3D positions and orientations. We got the shape of each object from the game's physics engine, which provides simple tight bounding polyhedrons as triangle meshes. The type associated with each object (e.g., cup, chair, or painting) was also provided by the game. We manually grouped some objects together into classes like book, food, clothing, and weapon to reduce the overall number of classes that need to be learned and to better fit the amount of data we have. We also removed a number of "atmospheric" or game artifact objects like cobwebs, ambient light sources, and hidden switches since they are either not placed by humans or relevant only to the game. The houses contain a total of 97 different objects to be classified though no house contained every type of object.

We chose 4 features sets to compare:

- Direction feature set
- Distance feature set
- Direction & Distance feature set
- Joint feature set

These feature sets use the proximity and direction features independently, together as separate features(e.g., X near Y, X left/below/inFront Y) , or joined together to form *joint* features (e.g., X near/left/below/inFront Y). The use of joint

TABLE I

FEATURE SET COMPARISON RESULT

Feature Set	Accuracy	Avg. Correct Rank
Direction	34.91%	7.85
Distance	37.17%	6.31
Direction & Distance	38.06%	6.07
Joint	43.97%	6.08

TABLE II

MAXIMUM ENTROPY/DECISION TREE COMPARISON RESULTS

Feature Set	Maxent Accuracy	Decision Tree Accuracy
Direction	34.91%	11.2%
Distance	37.17%	16.2%
Direction & Distance	38.06%	18.4%
Joint	43.97%	15.4%

features is experimentally justified if we assume that people use a common mental representation as the basis for their proximity and direction judgments [8]. In theory, these joint features should be strictly more useful than the simpler ones since they contain all the information of their component features. However, since there are significantly more possible types of relationships in joint features, it is more likely that there will be insufficient data to train with them.

The effectiveness of the feature sets was examined in three experiments. In each experiment we use a 10 fold cross validation test of each feature set's ability to classify objects. The specific classification task given was to identify object X given its shape and position and given information on the type, position and orientation of all the other objects Y in the house. This data is used to find qualitative spatial relationships between X and each Y which in turn were used for training the model and testing. In our first experiment, we examined each feature set's ability to perform object classification based on knowledge of the surrounding objects. In the second, we compared our maxent model to one trained using decision trees, a standard statistical classification technique. In the third, we examined how each feature set could handle varying degrees of simulated error in the training and test data.

The result of our maxent model given the spatial relationships of X with the surround objects Y is a list of probabilities of X for each known type of object. The two measures we used to determine the success rate of each classifier were accuracy and average rank. Accuracy was how often the object type with the highest probability returned by the classifier was the correct type. The average ranking was the average position of the correct type of the answer in the returned list of object type probabilities, sorted from most likely to least (i.e., a perfect system would have an average ranking of 1 since the first and most likely object returned would always be correct). We expected the accuracy rate to be low in comparison to known visual classifiers because exact object classification based on spatial relationships is impractical in many situations. The ranking result, however, demonstrates how effective the models are at approximating the correct type and how useful the model could be in reducing the difficulty of classifying objects using another classification approach.

A. Experiment 1: Feature Set Comparison

Table I shows the results of the first experiment, which compared each of the feature sets classification ability. Direction is clearly the least useful independent feature with larger differences in the feature sets ability to accurately determine object type and moderate differences in the average correct rank. One likely cause of this is that we are artificially imposing an internal frame of reference on each object to judge relative orientation and in many cases this is unrealistic and of little use. Distance is substantially more effective at both classification and average rank than direction. The combined distance & direction set was an improvement over the distance set, which indicates that there are situations were direction features can be used for identification when distance fails. Overall, we can see that increasing the amount of features in the training data leads to an improved classifier, an effect expected when using a maximum entropy model since they can handle large amounts of data and a wide variety of features.

The single most effective feature set was the joint features that combined both distance and directional features together. Joint features allow the classifier to identify more complex dependencies on spatial relationships between the objects such as above and touching which could not be created using the distance and direction features independently. Our original concern with this feature set was that it might be too fine grained and there would not be enough instances of each feature for the model to be trained effectively given our quantity of data. This, however, does not appear to be a problem given our large data source.

B. Experiment 2: Maximum Entropy/Decision Tree Comparison

In the second experiment we wanted to compare our maximum entropy classifier against decision trees, a standard statistical classification technique. The decision tree implementation we employed was C4.5, a commonly used technique which produces and then regularizes (prunes) the decision tree [20]. For each feature set, our maxent model and a decision tree were compared using 10 fold cross validation on the same data set. Our results are shown in Table II.

Clearly, maximum entropy models were the more effective classifier by a substantial margin. We believe that this is because, even after pruning, the decision trees were overfitted to the data. The resulting decision trees were very large, with approximately as many decisions as there were training examples. Their accuracy on the training data was frequently in the $70\% - 80\%$ range, much higher than could reasonably be expected, an indication that they might be overfitted.

C. Experiment 3: Maximum Entropy Model with Misclassification

Our third experiment examines how well our maxent classifier handles misclassified prior data since this will give us an indication of what level of precision in prior object classification is necessary for spatial object classification to be useful. In this experiment we simulated misclassification of the surrounding objects in both the training and test phases. Each experiment had a fixed n which is a percentage of misclassified object types in both the training and test phases. The type of a misclassified object was uniformly randomly sampled from the types of the other objects in the same environment. We used 10 fold cross validation and plotted the accuracy and average correct rank of each feature set to demonstrate how each handled prior object type misclassification.

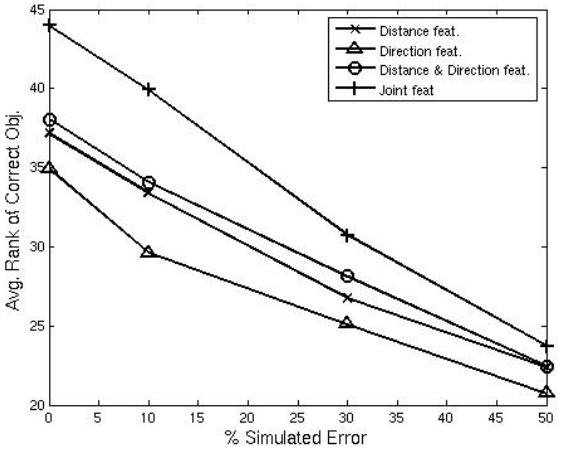

Fig. 3. A graph comparing the accuracy for each feature set. It was trained with our maximum entropy classifier under simulated training and test data object misclassification.

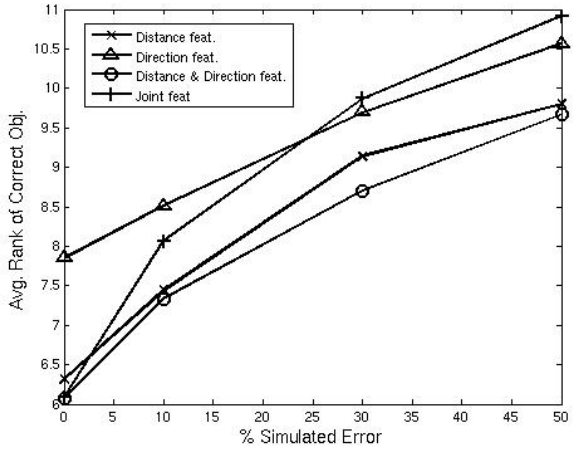

Fig. 4. A graph comparing the average rank of the correct object type in the returned results for each feature set. It was trained with our maximum entropy classifier under simulated training and test data object misclassification.

Figs. 3 and 4 show how the accuracy and average correct ranking degraded for each feature set under different amounts of simulated error. The accuracy test demonstrated that all feature sets accuracy appeared to degrade in a similar manner and at a similar rate. Direction and distance both seem to have a slightly faster rate of failure with error. We hypothesize that the cause of this is the greater number of relevant spatial relationships that can be learned with the large feature sets distance & direction and joint.

The average rank results, however, are quite different. The most noticeable effect here is that the joint feature set's average correct ranking degrades more rapidly than any of the other sets, performing substantially worse than all the others at a 50% error rate. The combined distance & direction set fared the best under error. The joint set seems more fragile in the presence of error, possibly because its features are more complex and it requires a greater amount of accurate data to find relevant relationships. The combined distance & direction set, however, can benefit from simple spatial relationships found with either distance or direction.

Given these results, it would appear that which feature set should be used depends on the given task. If the goal is classification accuracy, then clearly joint features are superior. However, if the goal is to give the correct solution the best average rank, then the combined set of distance & direction features is most useful and reliable under error. In the presence of misclassification, accuracy falls relatively quickly but the average ranking of objects can still be good at high error levels in the training and test data.

VI. Future Work

The work shown here is a preliminary step to demonstrate that qualitative spatial features can be used as the basis for learning object positional relationships. There are several improvements that are under development. One of the first will be an improved metric to determine proximity with regards to interposing objects. At present, we determine proximity by finding the minimum distance between the two objects, without reference to any interposing objects. This is problematic since there clearly is a qualitative difference between two objects being near with nothing interposed and objects being near with a wall separating them. Even with a technique that can detect an interposing object, we will still need to identify a feature language for describing the degree to which the path between the objects is obstructed.

Another area that we expect to see lead to substantial improvements is using a hierarchical classification approach to learning. For this, we first need to define for each object a hierarchical class structure (e.g., $cheddar \rightarrow cheese \rightarrow food \rightarrow object$). This could either be done manually or using data from an ontological data source like WordNet [13]. Then the problem becomes finding a way of training our classifier to learn multiple hierarchical class labels, though there has been some work on similar problems [27]. The result of this hierarchical approach would hopefully be both a more accurate classifier and the ability to determine an objects entire hierarchical type structure.

Finally, we are also very interested in other uses of our model than for object classification. In particular, we would like to be able to use it to solve the problem of determining the most probable position of an object in an environment. This relates back to one of the problems we described in our introduction, that of training a robot to be able to determine where in an environment it should be searching for a particular object. To do this, we need to reverse our model to determine the probability of a object being at a particular location in space and then sampling over the environment to find the most likely positions. This technique could also be applied to the generation of realistic virtual environments for commercial games. At present, it is a highly labor intensive task to construct the many locations in a game. It would be much simpler if the designers were able to build a few houses and then train a model to produce new environments based on those houses object structure.

VII. Conclusion

In conclusion, we have shown that some of the spatial structure of objects in human environments can be modeled using maximum entropy models and features based on the object's relative qualitative spatial relationships. Specifically, we demonstrated a spatial classifier based on a maximum entropy model and trained using qualitative spatial feature sets. The most successful feature set for classification accuracy was the joint features that were a combination of qualitative direction and distance features, though it had some problems at high error rates. The most successful feature set for giving the correct answer the highest average probabilistic ranking was a combination of our distance and direction feature sets that kept them as separate features. This feature set was more tolerant to error since it could identify simple relations from the distance and direction feature sets with less data than the joint set. We have also demonstrated a novel source of synthetic data on the layout of human environments from the game Elder Scrolls 4: Oblivion.

References

[1] S. Agarwal and A. Awan. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(11):1475–1490, 2004.

[2] M. Bar. Visual objects in context. *Nature Reviews: Neuroscience*, 5:617–629, August 2004.

[3] Bethesda Softworks. *Elder Scrolls 4: Oblivion*, 2006.

[4] M. Boutell, J. Luo, and C. Brown. Learning spatial configuration models using modified dirichlet priors. In *Proceedings of the 2004 Workshop on Statistical Relational Learning (in conjunction with ICML2004)*, 2004.

[5] A. G. Cohn and S. M. Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46(1-2):1–29, 2001.

[6] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43:1470–1480, 1972.

[7] G. Dorkó and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 634, Washington, DC, USA, 2003. IEEE Computer Society.

[8] M. Duckham, J. Lingham, K. T. Mason, and M. F. Worboys. Qualitative reasoning about consistency in geographic information. *Inf. Sci.*, 176(6):601–627, 2006.

[9] L. Fei-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, Washington, DC, USA, 2003. IEEE Computer Society.

[10] J. Freeman. The modeling of spatial relations. *Computer Graphics and Image Processing*, 4:156–171, 1975.

[11] S. Gottschalk, M. C. Lin, and D. Manocha. OBBTree: A hierarchical structure for rapid interference detection. *Computer Graphics*, 30(Annual Conference Series):171–180, 1996.

[12] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957.

[13] J.-U. Kietz, A. Maedche, and R. Volz. A method for semi-automatic ontology acquisition from a corporate intranet. In *Proceedings of EKAW-2000 Workshop "Ontologies and Text", Juan-Les-Pins, France, October 2000*, number 1937 in Springer Lecture Notes in Artificial Intelligence (LNAI), 2000.

[14] S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1150, Washington, DC, USA, 2003. IEEE Computer Society.

[15] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV '05)*, volume 2, pages 1284 – 1291, October 2005.

[16] K. Murphy, A. Torralba, and W. T. Freeman. Graphical model for recognizing scenes and objects. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[17] B. Neumann and R. Möller. On scene interpretation with description logics. *Image and Vision Computing, Special Issue on Cognitive Vision*, 2007. to appear.

[18] S. J. Phillips, M. Dudik, and R. E. Schapire. A maximum entropy approach to species distribution modeling. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 83, New York, NY, USA, 2004. ACM Press.

[19] S. D. Pietra, V. D. Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(4):380–393, 1997.

[20] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[21] A. Ranganathan and F. Dellaert. Semantic modeling of places using objects. In *Proceedings of Robotics: Science and Systems*, Atlanta, GA, USA, June 2007.

[22] A. Ratnaparkhi. A simple introduction to maximum entropy models for natural language processing. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania, 1997.

[23] R. D. Rimey. Control of selective perception using bayes nets and decision theory. Technical Report TR468, University of Rochester, 1993.

[24] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *MIT AI Lab Memo AIM-2005-025*, 2005.

[25] J. Vogel and K. Murphy. A non-myopic approach to visual search. In *CRV '07: Proceedings of the Fourth Canadian Conference on Computer and Robot Vision*, pages 227–234, Washington, DC, USA, 2007. IEEE Computer Society.

[26] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proceedings of the European Conference on Computer Vision, 2000*, pages 18–32, 2000.

[27] S. Zhu, X. Ji, W. Xu, and Y. Gong. Multi-labelled classification using maximum entropy method. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 274–281, New York, NY, USA, 2005. ACM Press.