Minimally-Supervised Classification using Multiple Observation Sets

Chris Stauffer
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

This paper discusses building complex classifiers from a single labeled example and vast number of unlabeled observation sets, each derived from observation of a single process or object. When data can be measured by observation, it is often plentiful and it is often possible to make more than one observation of the state of a process or object. This paper discusses how to exploit the variability across such sets of observations of the same object to estimate class labels for unlabeled examples given a minimal number of labeled examples. In contrast to similar semi-supervised classification procedures that define the likelihood that two observations share a label as a function of the embedded distance between the two observations, this method uses the Naive Bayes estimate of how often the two observations **did** *result from the same observed process. Exploiting this additional source of information in an iterative estimation procedure can generalize complex classification models from single labeled observations. Some examples involving classification of tracked objects in a low-dimensional feature space given thousands of unlabeled observation sets are used to illustrate the effectiveness of this method.*

1. Introduction

Current computational classification systems rely on excessively large numbers of labeled examples to classify effectively. In contrast, humans can classify objects extraordinarily well with very little supervision. While by no means offered as a complete explanation, it is interesting to note that humans almost never make a single observation of an object in the world (except arguably in cognitive science experiments). They are almost always able to see an object over an interval of time (even if it is a photograph). While the instantaneous change (e.g., optical flow) may be a major source of information for classification, it is our contention that exploiting the independent appearances or feature values can result in effective appearance-based classification of sequences as well as individual observations.

Whenever a process or object can be sampled multiple times, multiple observations of the *same* underlying process or object can be acquired. A Multiple Observation Set (MOS) is such a set of observations. The variability in visual observations from an MOS of a particular object may result from: noise in the sensors; active exploration (squinting one's eyes or moving a photograph closer to you); change in viewing angle; change in object position; change in object articulation; or any other change in the object state or sensor state that alters the observation that is made of the object.

At the very least, observations in a particular MOS should exhibit the type of noise incurred in estimating measurements of an object (e.g., camera noise). In some cases, the observations in an MOS exhibit class-conditional variability exhibited as a process evolves over time, e.g., a person's appearance during a walking cycle. In rare cases, the observations in an MOS can be considered to produce completely independent and identically distributed samples from the underlying class's appearance model.

This paper discusses a minimally-supervised learning system that can estimate complex, density-based classification models, often with just a single example per class. It accomplishes this by leveraging the information available in large quantities of MOSs to create effective density-based classification models. Examples of classification of tracked objects are explored.

1.1. Previous Work

Many researchers have induced a pair-wise similarity[1] measure from data that was originally embedded in a Euclidean space. Examples of these are exponential distributions based on Manhattan distance between observations [5] or squared distance [3, 6]. In the case of label assignment using a Markov random walk procedure, the underlying hypothesis is that these similarities are related to the likelihood that two observations share a label[2].

This paper leverages another source of information of which observations are likely to result from the same un-

[1] The similarity is often referred to as a distance or dissimilarity, but the underlying information is the same.

[2] Though this is an iterative, probabilistic technique, it is related to spectral clustering methods.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

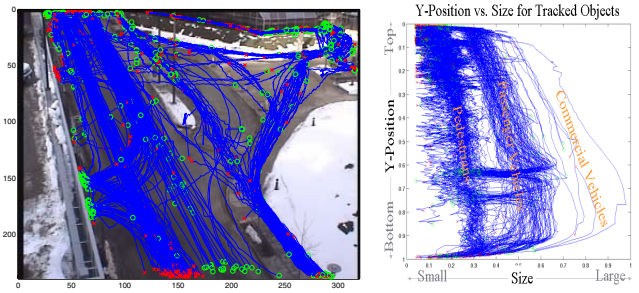
derlying class. Specifically, it leverages which observations *were historically* likely to result from the same underlying class. This paper will illustrate that this new source of information is vastly superior to simple local similarity estimates, particularly for sufficiently complex classification problems where data are not embedded in separable manifolds.

While this additional information is not universally available, there are many examples of where it is available or could be made available but is not exploited. Visual tracking is one example of where this type of data is plentiful and cheap. By tracking a single object, the variability in different features produced in observations of that object can be characterized. Because our research group has tracked millions of objects over the last seven years in indoor and outdoor environments, this is the primary area of application we will use to illustrate this method. This type of information has been used in unsupervised classification [4] to estimate the K hidden classes that generated the data. This was done using an EM procedure similar to Hofmann et al.[1] and Lee and Seung[2] to estimate the latent class-conditional, independent marginal distributions that best approximate the observed joint co-occurrences.

This paper centers on minimally-supervised and active learning, but exploits the same source of information. Szummer et al.[5] estimated class label likelihoods by estimating a K-neighbor connected graph with weights estimated based on a Gaussian distribution on distance and using said graph to propagate class label likelihoods for T time steps. This enabled effective classifiers to be built for many classic problems (e.g., the swiss roll) using a single example, but it has difficulty in less "classic" problems and requires three parameters to be determined (K neighbors, T time steps, and the variance of the noise process). Tishby and Slonim [6] estimated the probability of a Markov walk from each node ending at each other node. They found the number of time steps in which the mutual information of the complete NxN conditional distribution decreased the least and clustered the conditional distributions for each node at that time step.

In addition to using the MOS-based estimate that two observations share a label, our approach estimates the probability of a Markov walk from each labeled observation in a class to each unlabeled observation with a ρ-probability of restart. This procedure requires only a single parameter (the restart probability) and results in robust class likelihood estimation. By exploiting all of the MOS data to estimate transition likelihoods that are class-based rather than dependent on the embedding space, it is also independent of the choice of the embedding space (e.g., scaling axes, using radius vs. area, using velocity/direction vs. using dx/dy, etc.).

This paper begins by introducing the concept of MOSs. Section 3 describes the procedure that estimate class-conditional densities fromas few as a single observation. Section 4 discusses different supervision paradigms including: random labeling, informative labeling, corrective labeling, and elicited responses. Section 5 covers future research directions that relate to this learning paradigm. Section 6 draws conclusions from this work.



(a) (b)

Figure 1: This figure shows positional tracking data overlaid onto the scene from which it was captured (a) and the same data plotted on y-position vs. size axes (b). Each track is displayed as a blue line which begins at the origin state (green circle) and ends at its destination state (red x).

2. Multiple Observation Sets (MOSs)

An MOS is a set of observations of the *same* object or process in the world. Some examples of multiple observation sets are: multiple roles of a biased multi-sided die; the aspect ratio of a tracked object over time; the position of a tracked object over time; multiple independent answers to "What is your favorite number from 1-100?"; etc. In the case of tracking data[3], each tracking sequence results in a single MOS where the measured features of the tracked object at each time step are the samples and the MOS is the unordered set of those samples. For a set of features $f_i^1(t), f_i^2(t), ...$ for the i^{th} tracking sequence:

$$O_i = \{\{f_i^1(1), f_i^2(1), ...\}, \{f_i^1(2), f_i^2(2), ...\}, ..., \{f_i^1(T_i), f_i^2(T_i), ...\}\} \quad (1)$$

2.1. Discrete-output MOSs

Both time and the values of the features could be continuous. In practice, as a result of most sampling procedures, samples are only available at discrete points in time. Features can be either discrete (e.g., male or female) or continuous (e.g., position). Little can be said about continuous features unless some locality assumption is made. For example, given any number of labeled observations in a continuous space, a new observation has a zero probability of

[3] In the case of visual tracking, it is useful to tune a tracker to decrease the likelihood of falsely associating two objects

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

being *exactly* the same as any previous example and could be of a different class than its immediate neighbors given a sufficiently complex classification model.

Our assumption is that within a ϵ-area around a point in the continuous observation space, all observations have approximately equal likelihood of being produced by the same class. Under this assumption, we quantize the continuous space of observations into a set of discrete observations corresponding to ϵ-sized regions of the observation space. In low-dimensional observation spaces, this can be done by simply partitioning the space uniformly. In high-dimensional, sparse observation spaces this can be done by data-dependent quantization of the space. For simplicity, this publication will use uniformly placed bins of observations.

Figure 1(a) shows some positional tracking observations overlaid onto the scene from which they were captured over the period of an hour on a Friday afternoon between 3 PM and 4 PM. Figure 1(b) shows measurements of y-position and size for the corresponding data. The scene consists of: a roadway; a circular driveway; and sidewalks. Tracking sequences of less than two seconds in duration have been removed. What remains are 586 tracked objects, mostly vehicles and pedestrians.

Many of the characteristics of the data are expected. For example, objects are larger if they are in the foreground of the camera (further down in the camera view). Also, the estimated size of pedestrians is very noisy relative to the size of vehicles. Those less familiar with visual tracking data may be surprised by the decrease in object size at certain locations. This results from objects undergoing occlusion as they enter or exit the scene. This is one of the factors that makes classification of realistic tracking data a non-trivial problem.

3. Iterative estimation procedure

This section describes the iterative procedure that estimates complex, class-conditional densities from a single labeled observation. The subsections describe estimation of the data-dependent transition likelihoods, the Markov random walk estimation procedure, derivation of the class-conditional likelihoods using a Markov walk with probabilistic restart, and classification of novel observations and observation sequences. The next section will describe some applications of these classifiers.

3.1. Transition likelihoods

Previous machine learning algorithms have been used to estimate labels for unlabeled observations from the labels of their "neighbors". Given no additional information, one is forced to estimate the "neighborhood" relationships for pairs of points from their physical relationship in the embedding space. In previous work, these estimates of similarity have been functions of distance between data points (sometimes set to zero for data points that are not the first K-neighbors). Thus, the transition likelihoods were simply a function of the coordinates of the original embedded observations.

The stated goal of many of these approaches was to propagate labeling information along high-density separable data "manifolds." Unfortunately for many applications, object classes are not at all separable. In fact, in the experience of this author, the areas of the input space with the highest density often correspond to ambiguous observations that could be produced by multiple classes. These areas result in significant "bleeding" of class-conditional density estimates into un-related classes.

As an example, Szummer and Jaakkola [5] used

$$p_{ik} = \frac{W_{ik}}{\sum_j W_{ij}} \tag{2}$$

where W_{ik} is defined as

$$W_{ij} = \exp\left(-d(x_i, x_j)/\sigma\right). \tag{3}$$

where $d(x_i, x_k)$ is any valid distance function.

In contrast, we define the transition likelihoods as the probability that an x_k observation resulted from observing an object that also produced an x_i observation. We define $\mathbf{\Phi}$ as the set of all pairs of observations in the entire set of MOSs (excluding pairing observations in an MOS with themselves). $\mathbf{\Phi_i}$ is the subsets which contain the observation x_i as the first element and $\mathbf{\Phi_{ik}}$ is the subset which contain $\{x_i, x_k\}$ pairs. The transition likelihood is simply the likelihood of drawing an $\{x_i, x_k\}$ pair from $\mathbf{\Phi_i}$.

$$p_{ik} = p(\mathbf{\Phi_{ik}}|\mathbf{\Phi_i}) \tag{4}$$

$$\tag{5}$$

This value is low for two observations that are rarely exhibited by the same class and high for two observations that are often exhibited by the same class.

3.2. Markov Random Walk

For class l, with N labeled observations $\{o_1, o_2, ..., o_N\}$, the likelihood of each observation given just the observation set is the likelihood of each observation given a random draw from the labeled examples for that class, or

$$p_0(x_i|c_i = l) \equiv \frac{\sum_{n=1}^{N} \delta_{o_n, x_i}}{N} \tag{6}$$

where $\delta_{o_n, x_i} = 1$ if the n^{th} observation is equal to x_i. As $N \rightarrow \inf$, this estimate approaches the true class-conditional density. But given a single observation per

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

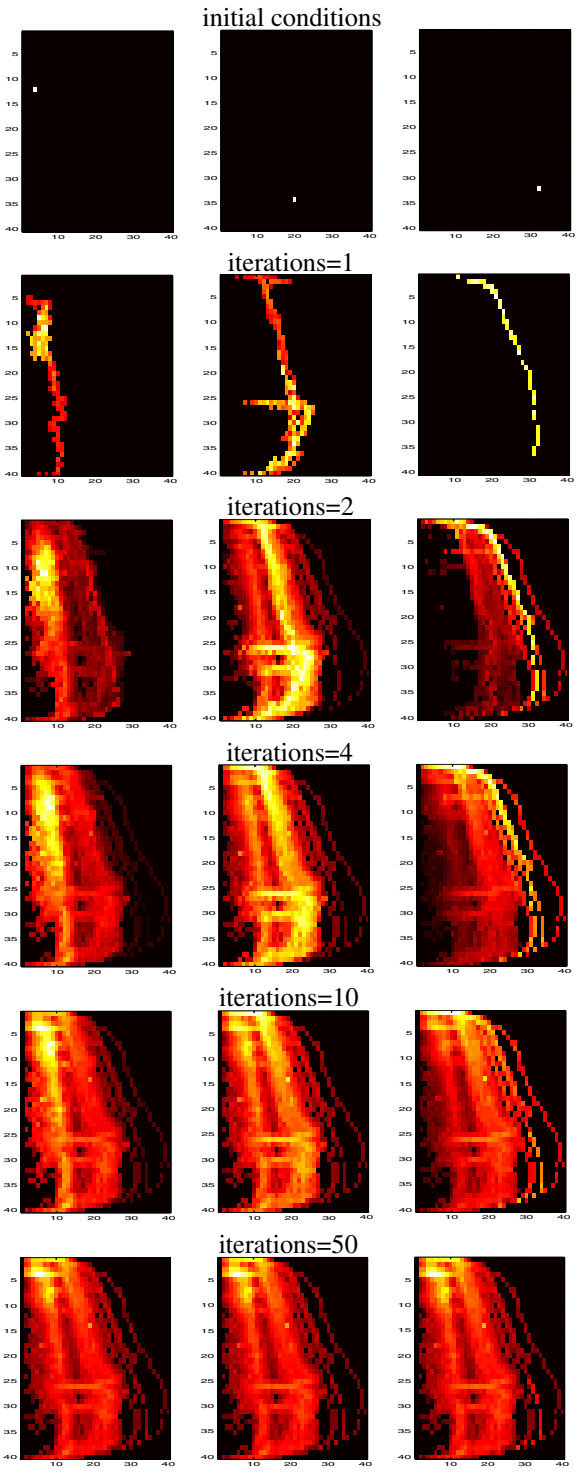


Figure 2: This figure shows the likelihood of each observation given all possible random walks of length 0, 1, 2, 4, 10, and 50 steps for three initial conditions. The initial conditions (top row) correspond to one labeled observation per class– pedestrians, passenger vehicles, and commercial vehicles respectively. The observation space is y-pos (vertical axis) and size (horizontal axis). Brighter values are more likely.

class, this density might look similar to the first row of Figure 2.

The class-conditional densities are estimated by assuming that an observations likelihood can be inferred based on its likelihood and the likelihood of other observations that have been exhibited by the same class. This recursive definition can be formulated as a first-order Markov random walk where each belief state can be inferred from the belief state at the prior time step. Thus, the likelihood of a random walk of t iterations is defined recursively as follows

$$p_t(x_i|c=l) = \sum_j p_{ij} p_{t-1}(x_j|c=l). \tag{7}$$

Figure 2 shows the likelihood of this Markov random walk for three initial conditions over increasing numbers of iterations. These three initial conditions correspond to single, labeled observations (y-position and size) of a pedestrian, a passenger vehicle, and a commercial vehicle. The first row shows that the likelihood is one for the labeled example for each class and zero elsewhere before any iterations. The second row shows $p(\mathbf{\Phi_{ij}}|\mathbf{\Phi_i})$ for the three labeled values of x_i. This represents the distribution of x_j observations that occurred in the same MOSs as the three labeled observations. As the number of iterations increase, the densities generalize to observations that have no direct relationship to any labeled observations. As $t \rightarrow \inf$, the all three Markov processes converge to a stationary distribution, which is independent of initial conditions.

To use this type of estimation a number of time steps must be chosen. If this number is too small, some number of observations may have no likelihood under any label. If this number is too large, the densities will be largely independent of their initial conditions.

3.3. Markov Walk with Restart

Rather than choose a specific number of iterations to estimate the class-conditional likelihood densities, we estimate the likelihood of each observation given an infinite random walk with restart probability, ρ.

$$\begin{aligned} p_t(x_i|c=l) &= (1-\rho) * \sum_j p(x_i|x_j, \mathbf{\Phi}) p_{t-1}(x_j|c=l) \\ &\quad + \rho * p_0(x_i|c=l). \end{aligned} \tag{8}$$

Figure 3 shows the converged likelihoods for different values of ρ. For $\rho = 1$, only the labeled observation for each class has a non-zero probability. For extremely low values of ρ, the likelihood functions become increasing independent of the labeled data. But for most values between 0.1 and 1.0, the likelihood functions will result in effective classification using a single example for each class. For future sections, we will refer to the converged estimate (where $t \rightarrow \inf$) of the likelihood of a observation x_i for a particular value of ρ given a particular class l as $p_\rho(x_i|c=l)$.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

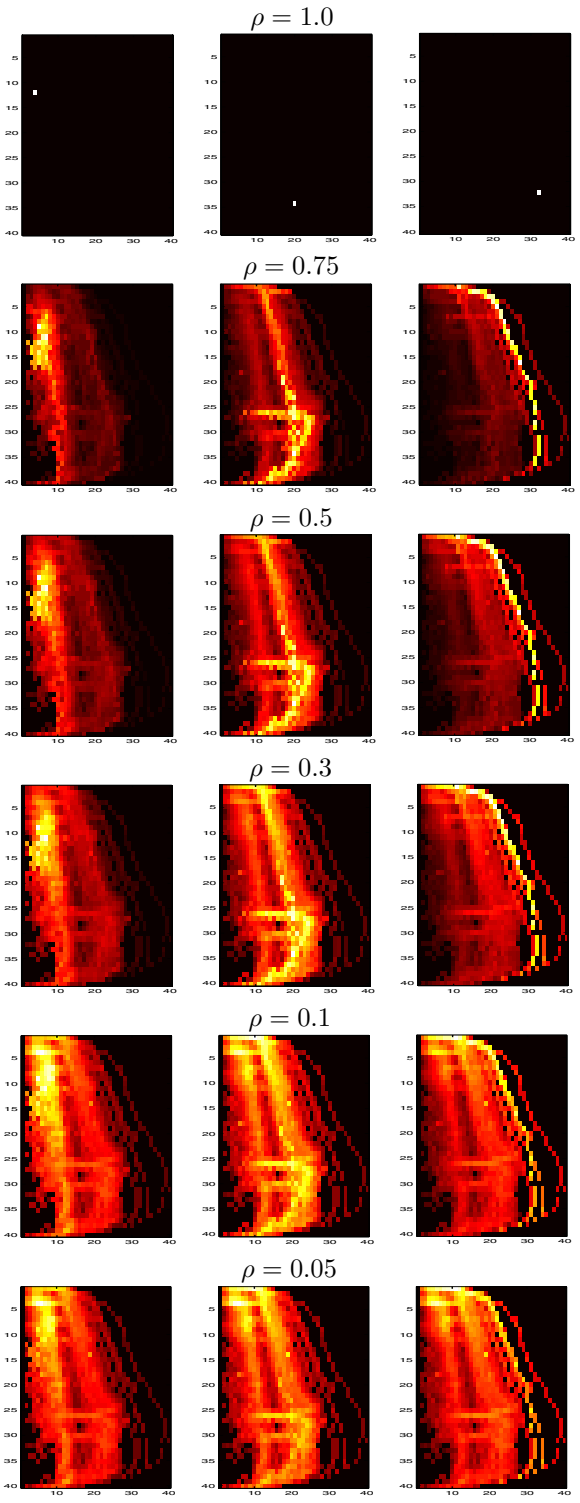


Figure 3: This figure shows the likelihood of each observation in $\{y - position, size\}$-space given three different labeled observations– a pedestrian; a passenger vehicle; and a commercial vehicle. Each row shows the density for the Markov walk with restart for $t \rightarrow \inf$ given different values of ρ.

3.4. Classification

The likelihood of an MOS, $O_i = \{x_1, x_2, ..., x_n\}$, under class model l is

$$p(O_i|c = l) = \prod_{j=1}^{n} p(c = l)p_\rho(x_j|c = l). \quad (9)$$

where $p(c = l)$ is the prior likelihood of observing a particular class l. If the exact value of this probability is known, it should be used. If it is not known, it can often be effectively estimated from the N original data points as

$$\hat{p}(c = l) = \frac{\sum_{i=1}^{N} p(x_i|c = l)}{N}. \quad (10)$$

This is a probabilistic estimate of the amount of data that is likely under each density.

Classifying MOSs with more than one element optimally in classification domains that are not separable requires a density-based classification. It is more effective than discriminant classification because it represents uncertainty in ambiguous observations. For instance, when any tracked object enters a scene, its size not discriminant, so it may be likely under multiple class models.

For $\rho = 0.75$ in Figure 3 three classes are shown derived from single examples of a pedestrian, a passenger vehicle, and a commercial vehicle. At the y-position where vehicles leave the parking ramp, vehicles are the size of a pedestrian. Once the car is unoccluded, it will have a high likelihood under the car model and a low likelihood under the pedestrian model. It is interesting that the commercial vehicle class does not have a high likelihood of occlusion from the parking garage. Thus, a vehicle leaving the parking garage is less likely to be a commercial vehicle.

4. Learning Paradigms

Thus far, we've discussed estimating class-conditional density models from a one (or a few) labeled observations. This section discusses different methods for choosing the examples and how those choices affect the amount of supervision necessary for a particular performance.

Classification results are shown for random labeling, informative labeling, corrective labeling, and elicited responses. Random and informative labeling involve classifying previously unseen observations using only a small set of labeled observations from each class, selected either randomly or pseudo-randomly. Corrective labeling and elicited response are two online learning paradigms. Through this section, it should become evident why density-based estimators are useful for these learning paradigms. Our results illustrate impressive generalization given minimal supervision.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

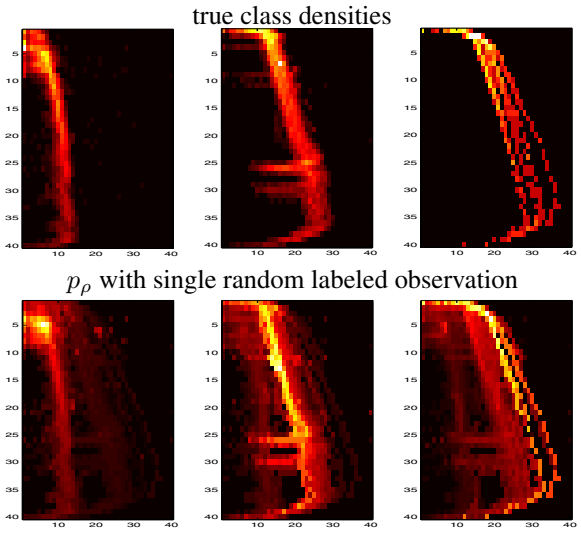


Figure 4: This figure shows the true class-conditional observation densities measured from a labeled corpus and the densities approximated from a single labeled example of each class (pedestrians, passenger vehicles, and commercial vehicles) for a given $\rho = .75$.

4.1. Random Labeling

By labeling randomly selected observations, a classifier can be quickly estimated. Figure 4 shows the true class-conditional observation densities (estimated from a labeled corpus) and the class-conditional observation densities approximated from a *single* random example of each of three classes– pedestrians, passenger vehicles, and commercial vehicles. The pedestrians (left) were primarily near the top of this scene and exhibited the smallest sizes overall. The commercial vehicles (right) were the largest vehicles and tended to pass through the entire scene. The passenger vehicles (middle) were moderate sized, except at the 4 locations where they entered or exited the scenes (the top, bottom, and two parking garage entrances in the lower third of the scene).

The observation densities resulting from a single labeled example are reasonably complex. They implicitly normalize object sizes because of the characteristics of objects as they move through the environment, not because of an explicit model of normalization. As a result they also effectively represent unusual data artifacts, like persistent scene occlusions.

Interestingly, a pedestrian-sized object near the parking garage exit is more likely to be a car than a pedestrian. But, because the likelihood under both classes is high, a pedestrian sequence that passes through this region of ambiguity is likely to be properly classified shortly after it leaves the area of ambiguity. This type of region of high density observations often results from ambiguous observations shared by multiple classes. This common (realistic) circumstance is where standard manifold label propagation algorithms run into difficulty. This location would result in significant "bleeding" between the pedestrian and vehicle classes with the manifold propagation techniques.

	Random	Informative
Single	$89.5 \pm .34\%$*	$93.5 \pm .15\%$*
Pair	$89.9 \pm .23\%$	$94.8 \pm .07\%$
Sequence	$91.0 \pm .79\%$	$95.4 \pm .17\%$

Table 1: This table shows classification results for a classifier trained with a single example per class. The rows correspond to classifying individual observations, pairs of observations from a single sequence, and entire sequences. The columns correspond to experiments with randomly selected examples and informatively selected examples. The mean and variance of performance is shown for 20 runs. Note(*): because the classes are not separable, the best performance attainable on this labeled data set is 96.86%

Some classification results for 423 objects are listed in Table 1. For all of our experiments, the unlabeled observations used to learn the propagation densities, the labeled observations, and the test sets were mutually exclusive. For a single labeled example per class, the performance is exceptional. The reason for the maximum performance of 96.86% for classification of single observations is that there are many pedestrians and vehicles near the top of the scene which are always occluded (i.e. the Bayes error rate is 3.14%). If those cases were discounted, the average classification performance for a single, randomly-selected example would have 92.4%. By using MOSs of length 2 or entire MOSs, the classification improves by a percent or two.

4.2. Informative Labeling

The reason for the high variance in the classifiers built from randomly chosen observations is that in some cases an ambiguous observation is chosen as an exemplar for a class. Choosing not to label ambiguous observations can significantly reduce the number of labeled observations required to achieve a defined level of performance. Table 1 illustrates that if a supervision source is able to pass up its first observation in favor of a second, less-ambiguous observation, the performance is significantly enhanced. Unfortunately, this requires some domain-specific knowledge on the part of the supervisor.

4.3. Corrective Labeling

If a supervision source is continuously available to monitor the performance of a system (e.g., a security guard), a classification system can be initialized with nothing more than

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

the knowledge of which classes are present in the scene.

As each object passes through the scene the system labels the objects. If the system labels an object incorrectly, the supervisor corrects the classification. E.g. "Daddy, [child points] there's an airplane!" "No son, that's a helicopter." Based on the presented examples, the class densities and prior probabilities can be re-estimated. As this process continues, the system will require less and less correction as the system approaches its maximum performance. Figure 6 shows some results for this type of supervision.

4.4. Elicited Responses

Unfortunately, in many environments this process could be very tedious because 98% of the objects may be of the same class showing the same type of variation. To solve this problem, we introduced an "unknown" class with an equal likelihood of producing any observation and an initial prior weight. Thus,

$$p(c = l_u|x_i) = \frac{p(c = l_u)P(x_i|c = l_u)}{\sum_{l_i \in labels} p(c = l_i)p(x_i|c = l_i)}. \quad (11)$$

This can be used in two ways. In the first method, an online system can ask for supervision only if the likelihood of the labeled example is higher under the "unknown" hypothesis than any of the existing class models. After the label is incorporated, that observation (and all similar observations) will have a higher likelihood under the the given class than the unknown class. This will significantly reduce the amount of supervision required.

In a second method, a batch system could evaluate all unlabeled observations over a period of time and select the one that is most likely to be informative and query the supervisor on the class of that observation. By informative, we mean an observation that has a high likelihood under the "unknown" class. This observation will correspond to one that has a low likelihood under all the class-conditional models. By sampling from the unlabeled data in proportion to $p(c = l_u|x_i)$, one is likely to chose examples that are less related to previously labeled examples.

Figure 5(a) shows the scene we have been using as an example. Figure 5(b) shows the $size$ vs. $y-position$ class-conditional densities for the three classes of objects (pedestrians, passenger vehicles, and commercial vehicles). The fourth density in the row shows the likelihood that a sample will be chosen as informative, given the current state of the class-conditional models. In this case, the next query to the operator will likely be a pedestrian-sized object lower in the scene. After that, the next query will likely be another pedestrian or a large vehicle.

Figure 5(c) shows the same tracking data in $\{x, y\}$-image coordinates. The class-conditional densities in this example also result from a single labeled observation per

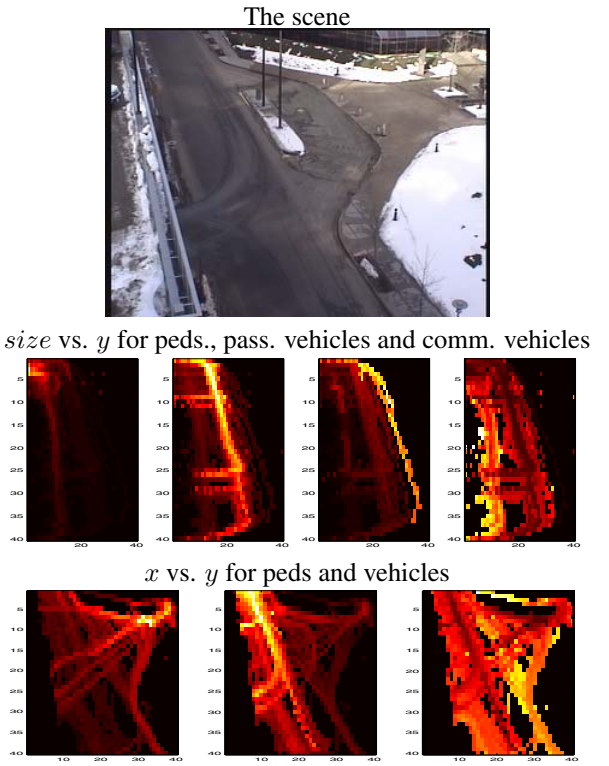


Figure 5: This figure shows the an image of the scene from which the data was taken. (b) shows the three class-conditional densities and the likelihood of drawing each observation from those that are informative. (c) shows the same information for a two class problem (pedestrians and vehicles) in {x,y}-space using the same data. Further discussion in the text.

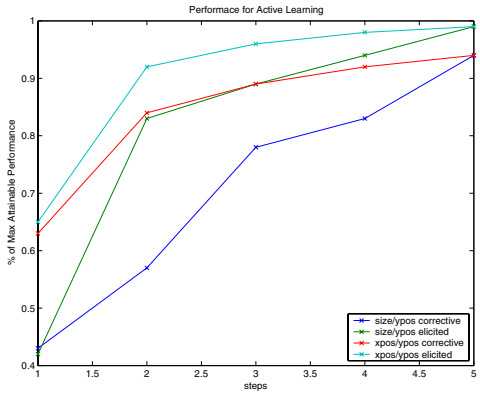


Figure 6: This figure shows the average classifier performance (relative the maximum performance) for the active learning scenarios.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

class. Given a single example of a pedestrian in the upper-right of the image, the class-conditional model shows that pedestrians are likely along many different paths from the building. From a single example of a vehicle leaving the scene, the class-conditional model generalized to have a high likelihood along the road in both directions and into and out of the parking garage. Without additional information, the classifier is 92% effective.

Two regions are shown to be particularly likely to be informative– the path near the building in the top-center and the sidewalk in the lower-right. The there were few examples of pedestrians on the path near the building that were tracked anywhere else in the scene, thus neither class is likely in this region and additional supervision is required. Any time there is no visual evidence that two clusters of observations might be related, additional supervision will be required to learn an effective classifier.

Figure 6 shows average performance of corrective and elicited response for the two classification problems described earlier over five time steps. It is apparent that choosing informative examples results in faster training, because the corrective labeling required more steps to achieve the same performance in both classification problems because many of the objects that were presented to the system were redundant with past observations. The elicited response system converged significantly faster in both cases. Note that the first problem was a three class problem and thus required *at least* three labeled observations to achieve reasonable performance. The second problem required at least two labeled observations.

5. Future Work

There are many areas for future investigation of methods that exploit Multiple Observation Sets. Many other descriptions of tracked objects could be added to the existing system including: velocity, direction, silhouette shape, component colors, mode of locomotion, etc. Other sources of data may contain similar information of the type of variation that should be expected within a class. E.g., biological data, user's actions, emails with the same subject heading.

The value for ρ was the same for all of our experiments and didn't have a substantial effect on classification performance, but the optimal value for ρ can depend on many factors including: the complexity of the observation space; the completeness of the MOSs; and the amount of supervision. One area of investigation is to adapt the value for ρ based on the amount of confusion in the observation space. As the amount of supervision increases, the value for ρ should increase. Obviously, in the extreme of infinite labeled data, ρ should be set to 1.0, as no generalization is required.

Though it is not remotely computationally feasible in the case of our tracking data. An alternative to the discretization step would be to use every continuous observation and propagate class labels only locally (using a kernel function similar to that used in previous approaches), but to estimate the likelihood of each observation using its entire MOS. Because of the discretization, our representation of p_{ik} remains constant size regardless of the number of MOSs.

6. Summary

This paper has presented a method for building robust classifiers with minimal supervision by exploiting Multiple Observation Sets (MOSs). It has outlined a method for estimating complex class-conditional densities from a single labeled observation. This method requires only a single parameter, ρ, and results in effective classification for reasonably difficult classification scenarios.

The density-based class-conditional model effectively represents uncertainty in ambiguous observations. MOSs with any number of observations can be classified using this model. This method lends itself to multiple models of supervision enabling classifiers to be quickly trained in novel spaces with minimal interaction. Though the results in classifying observations of tracked objects were especially promising, we believe this learning method could be applied to any domain in which equivalent observations sets are present.

References

[1] Thomas Hofmann. Probabilistic latent semantic analysis. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)*, 1999.

[2] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[3] Marina Meila and Jianbo Shi. A random walks view of spectral segmentation. In *Advances in Neural Information Processing Systems 13 (NIPS 2000)*, pages 873–879. MIT Press, 2000.

[4] Chris Stauffer and W. Eric L. Grimson. Automatic hierarchical classification using time-based co-occurrences. In *Computer Vision and Pattern Recognition*, pages 333–339, Fort Colins, CO, 1999.

[5] Martin Szummer and Tommi Jaakkola. Partially labeled classification with markov random walks. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

[6] N. Tishby and N. Slonim. Data clustering by markovian relazation and the information bottleneck method. In *Advances in Neural Information Processing Systems*, volume 13, 2000.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 **IEEE**