MODELING COUNT DATA FROM MULTIPLE SENSORS: A BUILDING OCCUPANCY MODEL

Jon Hutchins, Alexander Ihler, Padhraic Smyth

Department of Computer Science
University of California, Irvine
CA 92697–3425

ABSTRACT

Knowledge of the number of people in a building at a given time is crucial for applications such as emergency response. Sensors can be used to gather noisy measurements which when combined, can be used to make inferences about the location, movement and density of people. In this paper we describe a probabilistic model for predicting the occupancy of a building using networks of people-counting sensors. This model provides robust predictions given typical sensor noise as well as missing and corrupted data from malfunctioning sensors. We experimentally validate the model by comparing it to a baseline method using real data from a network of optical counting sensors in a campus building.

Index Terms— sensor networks, occupancy models, graphical models, Bayesian inference

1. INTRODUCTION

As sensors capable of monitoring daily human activity become increasingly affordable and ubiquitous, there is a corresponding need for algorithms capable of making sense of the resulting sensor observations across a wide variety of applications. One important subclass of such data are "count data," in which the observed signals consist of integer counts of the number of occurrences over time of a particular type of human activity. Examples include magnetic loop counters for monitoring freeway traffic [1], optical tripwires (or "people counters") for counting the number of people passing a particular point [2], and pre-processed video or optical motion detectors for monitoring a specific area [3].

Sensors that record count data often contain strong patterns reflecting the underlying rhythms of human activity. This periodic, predictable activity is referred to as "usual activity" in this paper. What makes these measurement streams complex, however, are random bursts of unusual or "event" activity, appearing as unusually high measurements (which can accompany a special seminar in a building or a baseball game in a stadium), or unusually low measurements (which might occur on a holiday).

In this paper, we extend earlier work on modeling count data at a single sensor [2] to a multi-sensor environment. A probabilistic model for each sensor, consisting of an inhomogeneous Poisson process for representing "usual" human activity and a hidden Markov process for representing bursts of unusual behavior. We describe how several such models can be coupled together to solve the occupancy problem for a building, namely, to infer an accurate estimate of how many people are in a building given noisy count data from its entrances and exits. The probabilistic nature of the model makes it relatively robust to both sensor noise and to sensor failure in the form of both missing and erroneous observations.

Obtaining accurate estimates of occupancy over time is an important component in many applications, including urban design and planning, security monitoring, and crisis response. For example, during a disaster crisis, information about the number of people and their locations is critical to first responders for allocation and deployment of resources.

The paper proceeds as follows. We describe the data and a simple baseline model in Section 2. In Section 3, we first review the probabilistic model for a single stream of count data, then show how individual sensor streams can be linked to form a multiple-sensor probabilistic model for building occupancy. Inference for the occupancy model follows in Section 4. Experimental evaluations to demonstrate the effectiveness of the model are described in Section 5, followed by conclusions in Section 6.

2. INFERRING OCCUPANCY FROM SENSOR DATA

Consider a trivial approach to occupancy estimation based on assuming that we have perfect information from a set of sensors about the number of people entering and exiting at each door in a building, i.e., no noise in the counts and complete coverage of all doors. Occupancy at time t is then simply the occupancy at time $t-1$, plus the sum (across sensors) of the counts of people who have entered since time $t-1$, minus the sum of counts of people who have exited.

Fig. 1 shows the result of estimating the building occupancy over a 5 week period using this trivial method. This graph is derived using data from optical "people counter" sensors that report aggregate counts every 30 minutes at 6 doors for a particular building (CalIT2 on the UCI campus). We immediately see from Fig. 1 that the simple approach produces very poor results, with a systematic negative trend in the estimate of the number of people in the building.

This problem arises because the sensors are imperfect, with noise corresponding to both under- and over-counting. The sensors used in Fig. 1 are pairs of optical sensors that register a count when an optical beam is interrupted. They are spaced in such a way as to determine whether a person is exiting or entering the building. "Nonhuman objects" can cause over-counts such as the one captured in the left panel of Fig. 2. More commonly, people entering in groups at the same time can cause under-counting such as is captured in the right panel of Fig. 2.

In addition, a sensor can fail outright. One of the largest discrepancies in Fig. 1 occurs at the beginning of week 5 and is partly due to a malfunction that occurred in a sensor at a door that is used more

This material is based upon work supported by the National Science Foundation under award numbers ITR-0331707, IIS-0083489 and IIS-0431085.

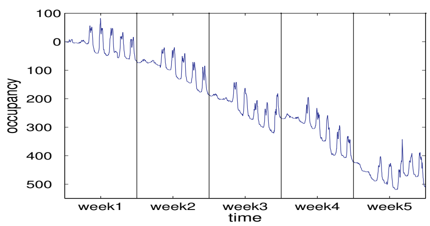


Fig. 1. An estimate of building occupancy assuming the measured values have no errors, so occupancy at time t equals that at time $t-1$, plus all incoming counts and minus all outgoing counts. Biases and miscounts can cause large systematic errors over a period of time.

Fig. 2. The left panel shows an example of a double count at the loading dock entrance of the building; the right panel shows an example of a missed count at the front door.

frequently used for incoming traffic than outgoing traffic. Like many systems, malfunctions for this type of sensor result in erroneous values, often zero, rather than any kind of explicit error signal.

One approach to counter the effects of the measurement noise is to simply enforce two constraints on the trivial estimation method above: (1) that occupancy can never be negative, and (2) that early every morning the building population should be zero. While incorporation of these types of constraints can improve the estimate quality, the results of this approach (which we refer to as the baseline method) are still quite inaccurate (see Section 5).

In the next section we outline a probabilistic model for the problem of estimating occupancy. This approach allows us to model sensor noise in a systematic manner, combine uncertain information from multiple sensors, leverage our prior beliefs about occupancy at particular times of the day, use statistical learning techniques to learn the parameters of our model from historical data, and systematically infer a probability distribution for occupancy over time conditioned on observed sensor data.

3. MODELING MULTI-SENSOR COUNT DATA

We use the framework of directed graphical models to capture relationships among different variables and parameters of interest. In this section we outline the structure of the model and in the following section we describe the inference process. Nodes in the graphical model represent random variables and probabilistic relationships are encoded as conditional distributions of child nodes given the values of their parent variables. The model contains unobserved (latent) variables representing quantities of interest (such as the true occupancy at time t) and parameters (such as Poisson rates); we are interested in reasoning about both conditioned on the observed evidence, i.e., the counts measured at sensors. Unless stated otherwise count variables such as N_t take non-negative integer values. The subscript t refers to a discrete time index, which for the count data used in

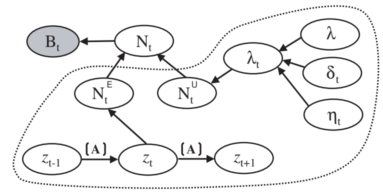


Fig. 3. Graphical model for a single stream of count measurements. Here, N_t represents the true number of counts at time t, B_t a noisy observation, N_t^U the counts due to "usual" activity (modeled by Poisson rate λ_t) and N_t^E any counts due to an event (modulated by the Markov process z_t).

this paper are spaced at half-hour intervals (the sensor report time), and a count such as N_t corresponds to an aggregate count over the half-hour prior to t.

Modeling a single sensor. We first describe a probabilistic model for a single sensor (whose graph is shown in Fig. 3); we then extend the model to multi-sensor data. Node B_t represents an observed count at time t for a particular sensor, a noisy version of the true (unobserved) count N_t for that sensor:

$$B_t = N_t + \Upsilon_t^O - \Upsilon_t^U \tag{1}$$

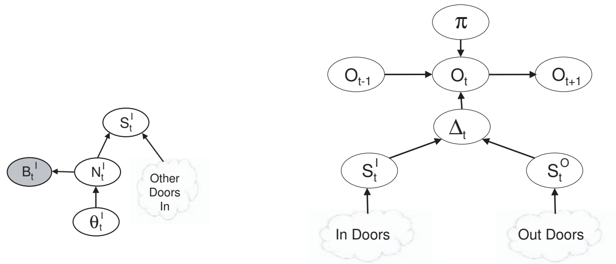
The number of undercounts Υ_t^U and overcounts Υ_t^O are modeled using separate binomial distributions: $\Upsilon_t^O \sim \mathrm{Bin}(B_t, v_O)$ and $\Upsilon_t^U \sim \mathrm{Bin}(N_t, v_U)$, subject to the constraint in Equation (1). This allows the the expected number of undercounts and overcounts to increase with the number of people using a door. In our experiments, we set $v_O = 1/70$ and $v_U = 1/20$ based on empirical observations of over- and undercounting.

The true count for a sensor, N_t, is modeled as the sum of two Poisson processes, where the two processes reflect usual activity for that sensor and bursts of abnormal activity ("events"): $N_t = N_t^U + N_t^E$.The component for usual activity, N_t^U, is modeled as a non-homogenous Poisson process. The event component N_t^E is an additional Poisson contribution governed by a Markov process, z_t, indicating whether or not an event is taking place at time t. z_t takes three values corresponding to an event with fewer people than normal ($z_t = -1$, e.g., a holiday), no event ($z_t = 0$), or an event with more people than normal ($z_t = 1$, e.g., a non-recurring large meeting in the building). $P(N_t^E|z_t)$ is Poisson with an unknown rate parameter for $z_t = -1$ and $z_t = 1$, and forces $N_t^E = 0$ for $z_t = 0$.

The non-homogenous Poisson process N_t^U depends on a time-varying rate parameter $\lambda_t = \lambda\delta_t\eta_t$, where the three components correspond to the average rate, λ, an adjustment for the day of week, δ_t, and an adjustment for the time of day, η_t (e.g., [4]).

For the event process, the Markov chain on z allows event persistence, which can lead to significantly better event detection performance compared to simpler threshold methods [2]. The transition matrix A defining the conditional distribution $P(z_t|z_{t-1})$ is also treated as a random variable in the graphical model. Except where stated, all prior distributions were chosen as in [2].

Inferring occupancy from multiple sensors. Each door to a building has separate data streams for the entrance ("in") counts and the exit ("out") counts. The true (unobserved) count for all of the in sensors at time t is represented by S_t^I, and similarly S_t^O for the out sensors. S_t^I and S_t^O are deterministic sums of the true count N_t for each in and out sensor, respectively. The occupancy at time t is



(a) (b)

Fig. 4. Linking the individual streams. (a) The sum node corresponding to the total building incoming (entry) flow. Θ represents the hidden parameters and variables specific to the individual stream, indicated by the dotted line in Fig. 3. (b) Graphical model for multiple-sensor building occupancy; total in and out traffic (S_t^I, S_t^O) modulates occupancy O_t.

denoted O_t, and is given by the sum of O_{t-1} and $\Delta_t = S_t^I - S_t^O$, which is the true (unobserved) change in occupancy over time-period t. These relationships are depicted in Fig. 4. For variables which take on countably infinitely many possible values (e.g., nonnegative integers) and for which no closed form exists for the conditional distributions of interest, we use heuristics to reduce the range of values under consideration.

We also include a geometric prior (with parameter set to .9 in the results in this paper) on O_t for $t = 3\text{AM}$, encouraging the model to leave few or no people in the building overnight. This helps to offset any systematic bias in the measurement noise which if unaccounted for could lead to ever-increasing or decreasing estimates of the number of people in the building (see Fig. 1).

4. INFERENCE

Given the probabilistic model described in the previous section, we now turn our attention to the inference problem, i.e., computing the conditional probability of quantities of interest (such as the occupancy O_t as a function of time) given both the observed measurements B_t (at all doors across all times of interest) and the priors. These quantities (the variables and parameters of the model) are learned by inferring their posterior distributions using Markov chain Monte Carlo (MCMC) sampling methods. In MCMC, we iteratively sample each set of variables given the current sampled values of the other variables in the model. After a sufficient number of iterations, these samples converge to the true posterior distribution.

Given a value of the true count N_t for each stream, we sample λ_t and z_t as described in [2]. Then, given both λ_t and z_t, we perform a forward–backward sampling procedure [5], similar to that used for z_t, to draw the total occupancy O_t and the true counts N_t for each sensor. In the forward inference pass, information flows from the individual streams up to the occupancy node and is combined with the belief about the occupancy found for the previous time slice. The backward pass then samples values for each of these variables. Since the graphical model is singly–connected given the λ_t and z_t, this procedure can be performed efficiently (in time linear in the number of measurements).

Let us define Λ to be the set of all λ_t for all streams and time, and Z similarly for z_t. Now, we compute the posterior distribution of O_t given Λ, Z, and the observed counts B_t. We first note that

$$p(N_t|\Lambda, Z, B_t) \propto p(B_t|N_t)p(N_t|\Lambda, Z)$$

by applying Bayes' rule and noting that B_t is conditionally independent given N_t. We can then compute the distribution of the variables S_t and Δ_t via successive convolution[1]. If we define the evidence E_t to be the set of all observations B_t at any of the sensors, this convolution process gives us the distribution $p(\Delta_t|E_t, \Lambda, Z)$.

The updated posterior of occupancy at time t is then

$$p(O_t|E_{1:t}) \propto \sum_{O_{t-1},\Delta_t} \delta(O_t - O_{t-1} - \Delta_t)\pi_t(O_t)p(O_{t-1}|E_{1:t-1})p(\Delta_t|E_t)$$

where $\delta(k) = 1$ for $k = 0$ and 0 otherwise (reflecting the deterministic relationship between O_t, O_{t-1}, and Δ_t), and where $\pi_t(O_t) = \text{Geom}(O_t; .9)$ when $t = 3\text{AM}$. We proceed forward in time to the maximum (or current) time $t = T$, then sample $O_T, O_{T-1}, \ldots$ backward to time $t = 1$. Given O_t and O_{t-1}, Δ_t is deterministic; the sum nodes S_t are sampled conditioned on their difference Δ_t, and the true counts N_t conditioned on their total S_t.

Given a sampled value for the true count N_t for each sensor stream, the sampling for the stream parameters Λ and the stream event process Z proceed as in [2]. Unlike [2], however, here the true count value N_t can change between iterations as the constraints of the occupancy model are enforced and as the belief about the true counts of the other sensor streams change.

5. EXPERIMENTS AND RESULTS

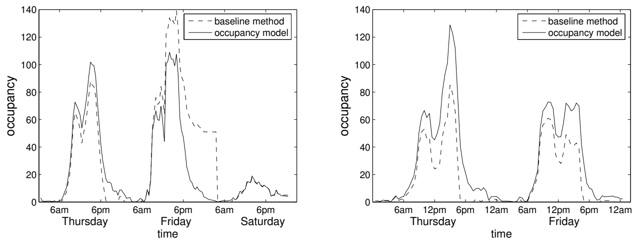
The data used in our experiments come from a campus building with six doors with optical people counters measuring the flow of people in both the entrance and exit directions. Nine weeks of measurements (6/11 to 8/12/2006) for each of the twelve streams were used for learning the model. All of the inference experiments in this section were run off-line, although on-line inference is a relatively straightforward extension of the techniques described in this paper.

In each of the experiments, the occupancy model is compared to a simple baseline method where two occupancy constraints are enforced: (a) occupancy can not be negative, so if $(O_{t-1} + \Delta_t) < 0$ then $O_t = 0$; and (b) occupancy is reset every 24 hours, so that at $t = 3\text{AM}$ we have $O_t = 0$.

Sensor Noise. The examples in this section contrast the occupancy model and the baseline method for days where the measured flow of people in entering or exiting the building is disproportionately larger than the flow in the opposite direction. This count difference is caused by the normal day-to-day noise of the sensor measurements.

Fig. 5(a) shows a day where using the baseline method would lead to the belief that approximately 50 people are in the building at 2:59 am. These 50 people are promptly forced to disappear via the 24 hour constraint. By smoothing, the occupancy model provides a more believable prediction for the day. Although we do not have ground truth, it is especially unlikely that the building held many people this particular Friday night since the two following days are weekend days with low activity and no large egresses.

[1] These operations are nominally $O(d^2)$ where d is the number of possible values entertained for each variable, but can be made $O(d \log d)$ via the fast Fourier transform [6].



(a) (b)

Fig. 5. (a) A day with more building entrance measurements than building exit measurements; the preceding and subsequent days are also shown for context. (b) Two days with more building exit measurements than building entrance measurements.

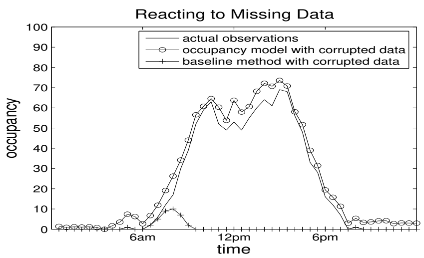


Fig. 6. The measurements for one building entrance stream accounting for approximately 50% of the total entrance counts were replaced by missing data labels.

Fig. 5(b) shows days with the opposite situation where more people are measured leaving the building than entering. The baseline model ignores all the extra exit counts at the end of the day, giving occupancy predictions that are likely too low. The probabilistic model, however, uses this information to adjust the occupancy at previous times upward, resulting in a more believable prediction.

Although we do not have a ground truth for comparison, these examples indicate that the probabilistic model provides more reasonable outputs than the baseline for typical amounts of sensor noise. In the next section, we investigate robustness to sensor failure in the form of missing or erroneous observations.

Validation. Since we do not have the true occupancy values, we address the issue of validating the model by removing some information and seeing how well the model recovers. We remove information in two ways: replacing observed measurements with missing labels, such as happens when a sensor stops communicating; and replacing observed measurements with corrupted data, as happens when a sensor malfunctions but continues to send false information.

In the first experiment, shown in Fig. 6, one day of measurements for the entrance stream of the main door to the building was replaced by missing labels. This particular door sees approximately 50% of the traffic to the building. The baseline method has no way to deal with missing data and degrades quickly. The occupancy model is able to recover much of the missing information, using the model of typical behavior (Poisson rate) and the information from the other streams such as extra exit counts.

Missing data is easier than corrupted data, since the model is alerted of the need to fill missing data with something reasonable. For corrupted data, we replace the measurements of one of the entrance streams with zeros, at a door used by roughly 25% of the building occupants. Two things help the multi-sensor model recover the missing information. First, the corrupted data appears unusual at the individual stream level, as the model expects data similar to the rate parameter. Second, if the corrupted data is only in one direction, the "excess" counts from the other direction will try to balance it out.

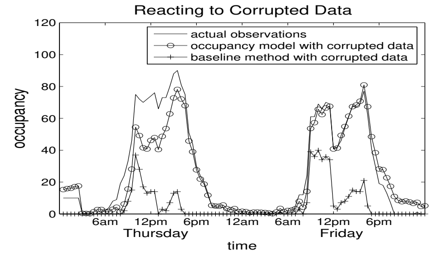


Fig. 7. Two days where the measurements from one entrance stream accounting for approximately 25% of the total entrance counts were replaced by zeros.

The results for corrupted data are shown in Fig. 7. As with the missing data experiment, the baseline method fails completely. The occupancy model performs much better, although it does not recover all of the missing information—the noise model resists deviation from the observed values, but the shared information is able to offset at least some effects of the corrupted data. This property of the occupancy model could also be used to detect a faulty sensor and provide an early alert prediction of a malfunction. A model with an explicit notion of sensor faults could improve performance still further.

6. CONCLUSIONS

Even with complete sensor coverage of all doors to a building, occupancy prediction is non-trivial. The probabilistic model presented in this paper overcomes many of the limitations of simpler methods. Spatial information and correlations among sensors will be examined in future work. In the long-term, we hope to combine people count sensors with other human activity sensing data such as electricity use, building schedules, and internet traffic to predict occupancy densities and future occupancy movements for larger areas such as a campus or city.

7. REFERENCES

[1] Freeway Performance Measurement System (PeMS), "`http://pems.eecs.berkeley.edu/`,".

[2] A. Ihler, J. Hutchins, and P. Smyth, "Adaptive event detection with time-varying Poisson processes," in *ACM Int'l Conf. Knowledge Discovery and Data mining*, 2006, pp. 207–216.

[3] C.R. Wren, D.C. Minnen, and S.G. Rao, "Similarity-based analysis for large networks of ultra-low resolution sensors," *Pattern Recognition*, vol. 39, no. 10, Oct. 2006.

[4] S. Scott, *Bayesian Methods and Extensions for the Two State Markov Modulated Poisson Process*, Ph.D. thesis, Harvard University, 1998.

[5] S. Scott and P. Smyth, "The Markov modulated Poisson process and Markov Poisson cascade with applications to web traffic data," *Bayesian Statistics*, vol. 7, pp. 671–680, 2003.

[6] P. Felzenszwalb and D. Huttenlocher, "Belief propagation for early vision," *Int'l J. of Comp. Vision*, vol. 70, no. 1, Oct. 2006.