Spatial-Temporal Understanding of Urban Scenes through Large Camera Network*

Jiejun Xu
University of California
Santa Barbara, USA
jiejun@cs.ucsb.edu

Zefeng Ni
University of California
Santa Barbara, USA
zefengni@ece.ucsb.edu

Carter De Leo
University of California
Santa Barbara, USA
cdeleo@ece.ucsb.edu

Thomas Kuo
University of California
Santa Barbara, USA
thekuo@ece.ucsb.edu

B. S. Manjunath
University of California
Santa Barbara, USA
manj@ece.ucsb.edu

ABSTRACT

Outdoor surveillance cameras have become prevalent as part of the urban infrastructure, and provided a good data source for studying urban dynamics. In this work, we provide a spatial-temporal analysis of 8 weeks of video data collected from the large outdoor camera network at UCSB campus, which consists of 27 cameras. We first apply simple vision algorithm to extract the crowdedness information in the scene. Then we further explore the relationship between the traffic pattern observed from the cameras with activities in the nearby area using additional knowledge such as campus class schedule. Finally we investigate the potential of discovering aggregated human movement pattern by assuming a simple probabilistic model. Experiment has shown promising results using the proposed method.

Categories and Subject Descriptors

H.1 [**MODELS AND PRINCIPLES**]: Miscellaneous

General Terms

Design, Measurement, Experimentation

Keywords

Sensing, Cross-modal Correlation, Camera Network

1. INTRODUCTION

There is a growing interest in studying and analyzing urban dynamics for purposes such as traffic forecasting, city planning and facility management. However, obtaining and modeling large, real world observational data is a challenging and costly task. In the past, researchers have used different digital traces of city-wide urban infrastructure as medium to study urban dynamics. For example, González et al. used cellular network data to study city dynamics and human mobility [5]. McNamara et al. [10] and Liu et al. [8] used data collected from RFID-enabled metro systems to monitor and predict co-location patterns among mass transit users. Froehlich et al. analyzed the digital footprint of bicycle usage from shared bicycling system to uncover patterns of human behavior and infer cultural and geographic aspects of the city [4]. All these methods utilize implicit data source to study urban dynamics.

On the other hand, outdoor cameras have become prevalent as part of the urban infrastructure. They provide an excellent data source to directly observe and understand urban dynamics. Previously, researchers have attempted to study videos from outdoor cameras, but they mainly focused on a few cameras covering small area and did not consider any relationship between visual data and external knowledge of the scene [6]. Here we demonstrate the potential of urban dynamic understanding through multiple cameras covering a large area. In particular, we collected videos from a network consisting of 27 cameras at the University of California at Santa Barbara (UCSB) [7] spanning a period of eight weeks. We show that better scene understanding can be achieved by analyzing the videos with simple vision techniques along with additional information of the area, such as campus class schedule. The main contributions of this paper are:

- Gaining insights of the urban scene through analyzing videos in a large camera network. To the best of our knowledge, this is the first attempt to understand the dynamic scenes in large area through a camera network over an extended period.
- Reducing uncertainty in visual observation through information fusion. By incorporating campus class schedules, we can better explore the relationship between the traffics observed from the cameras with activities in the nearby area.
- Demonstrating the potential of discovering aggregated human movement pattern. Assuming a simple probability model, we can obtain a fine-scale spatial-temporal estimation of the pattern.

*This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA) and ONR grant #N00014-10-1-0478

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MPVA'10, October 29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-4503-0167-1/10/10 ...$10.00.

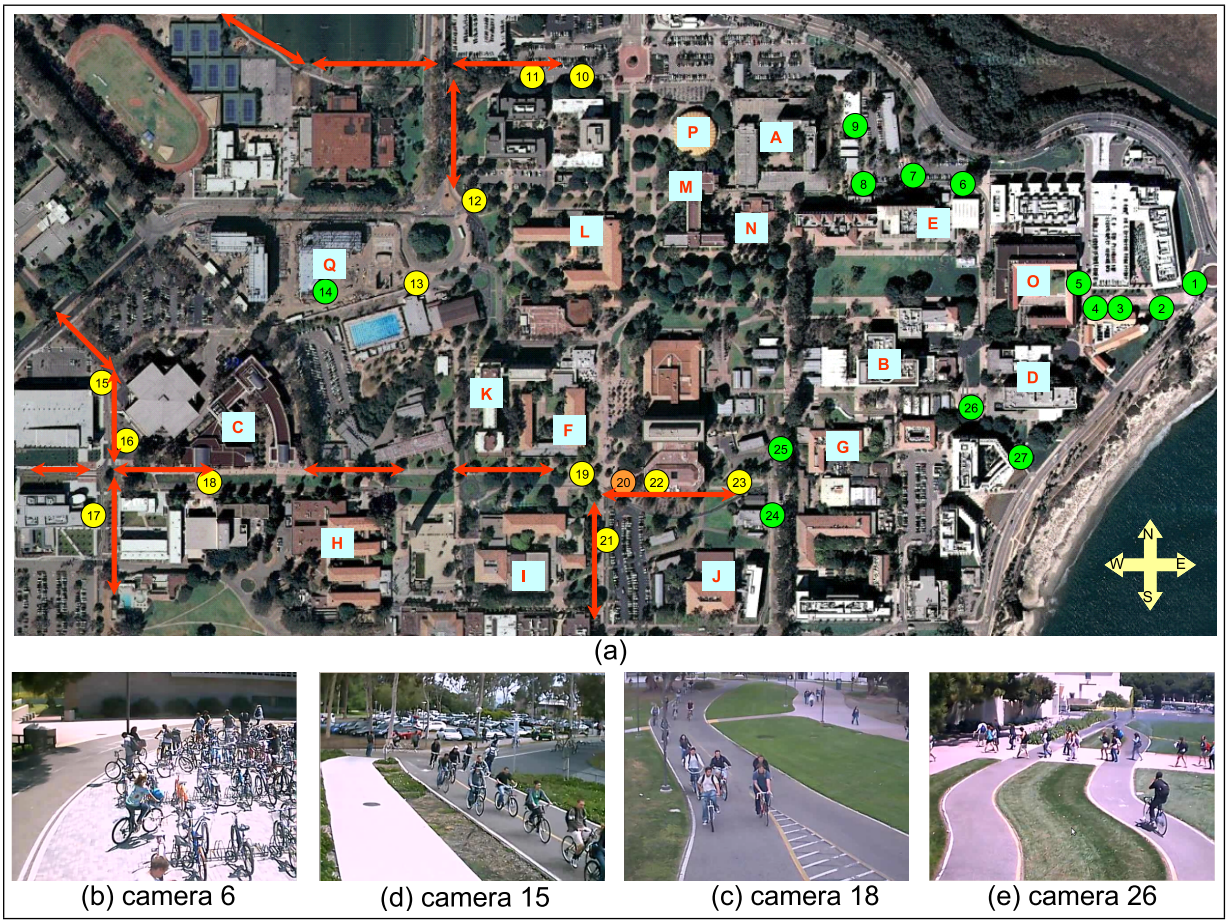

Figure 1: (a) UCSB campus: Squares with alphabets indicate buildings; Circles with numbers indicate camera locations; Color of the circle indicates camera category (Yellow: global camera; Green: local camera; Orange: global camera but wrongly classified as local camera. See section 3.2 for details). (b)-(e) snapshots of sample camera views. Major paths for non-motorized traffic are marked in red arrows.

The rest of the paper is organized as follows. Section 2 explains the camera network dataset for our experiment. Section 3 shows the details of our approach to analyze the videos to obtain spatial-temporal understanding of the scenes. Interesting findings from our reasoning will be shown in this section. Finally we will have our conclusion and future work at section 4.

2. THE UCSB CAMPUS DATASET

We collected our data from a recently deployed camera network at UCSB campus [7]. The network consists of over 40 stationary cameras covering wide area of the campus. In our experiments, we only use data from a subset of 27 cameras as show in Figure 1, due to the reason that many cameras are still under frequent testing. All of the 27 cameras are fixed outdoor cameras monitoring either campus bike paths or pedestrian walkways, i.e. observing non-motorized traffics. Since all the outdoor cameras in this network are powered by batteries, they operate only on discrete time periods. The recording time covers Monday to Friday, and each camera captures 200 minutes of video every day. The 200 minutes are spread to ten 20-minute recording windows centered on the hour from 8am to 5 pm (e.g. 7:50am-8:10am, 8:50am-9:10am, . . . , 4:50pm-5:10pm). In other words, there are 50 20-minute recordings V_h in a week for each camera ($h = 1, 2, \ldots, 50$). Our entire dataset spans a period of eight weeks (April and May) in the Spring quarter of 2010, and consists of more than 3000 hours of video in total.

3. SPATIAL-TEMPORAL REASONING

3.1 Visual Processing

Before we can explore the dynamics of the campus, we need to extract information from the videos to describe the scene. We chose to describe the scene by the level of crowdedness. In a typical vision setting, this is done in several steps, detecting people, tracking them over time, and counting the number of tracks. However, each of these steps is itself a challenging task, and they usually come with high computation cost and low accuracy. Motivated by the recent work of [2, 9], we argue that precise people counting might not be necessary to describe the scene. Instead, we propose to estimate the crowdedness of the scene by looking at the local motion within the video frames. This is because local motion are mostly caused by human movement given fixed camera setting. For each V_h, we extract the optical flow [1] in each frame. The clip is divided into shots of duration T_w mins (e.g., $T_w = 1/4 = 15$seconds). For each shot,

we then calculate the average optical flow, and define it as $As(T; c, h)$ (Activity Score at time $T = 0, 1, \ldots, \frac{20}{T_w} - 1$ in camera c centered at hour h). Given most of the motion within the camera view is caused by human movement, we think that the average optical flow within a short duration T_w serves as a good indication of how crowded a scene is.

Figure 2 shows a snapshot from camera 1 and its corresponding $As(T; 1, h)$ with $T_w = 20$ over a week, averaged across the eight-week observation period[1]. This camera is mounted next to a bus station. A repeating "5pm spike" appears across all weekday, which reveals the usual commute patterns for this station. The relative magnitude of the "5pm spike" is lower on Friday, which corresponds to the fact that many people leave earlier on Friday afternoon.

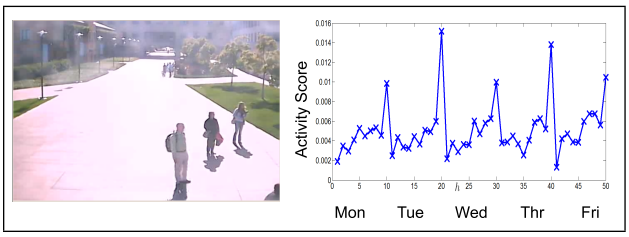

Figure 2: Camera 1 and its $As(T; 1, h)$ with $T_w = 20$.

Figure 3 shows the $As(T; c, h)$ of two other cameras with finer granularity ($T_w = 0.25 = 15sec$). We can easily observe that average activity is higher on Monday, Wednesday and Friday. In addition, both $As(T; 15, h)$ and $As(T; 18, h)$ decrease consistently in the afternoon, which also correspond to typical university traffic pattern. From the snapshot, we can see that $As(T; c, h)$ indeed reflects the crowdedness of the scene.

3.2 Camera-Building Correlation

In a campus scenario, most human movement (except for weekend/holidays) are caused by students attending classes at different buildings. Given this assumption, there should be a strong correlation between camera's observation $As(T; c, h)$ and class schedules. In UCSB campus, most classes start on the hour (e.g. 9am), which overlaps with our recording windows. For each building b at hour h, we define $N(b, h)$, which is the total number of students attending classes inside the building at that time. These information are available to the public through class schedules listed on university web page. In our experiments, we have obtained class schedules for 17 buildings, which contain most of the classrooms for lecturing. Squares with labels in Figure 1(a) show the locations of these buildings. The rest of the unlabeled buildings are mostly research laboratories and offices, which usually cause much less traffics/activies across campus. Figure 4 illustrates the schedule information for building **E**.

For each camera, one obvious question to ask is that, what are the buildings related to the activities observed by this camera. This question can be answered if we have camera calibration parameters, campus blueprint, and most importantly human mobility patterns (i.e., which route students usually take to reach a building). All these information is

[1]The camera network at UCSB [7] is still at the experimental stage, the duration of V_h varies. Also, student movement on campus exhibits a strong weekly temporal regularity, thus we simply average $As(T; h, c)$ over eight weeks.

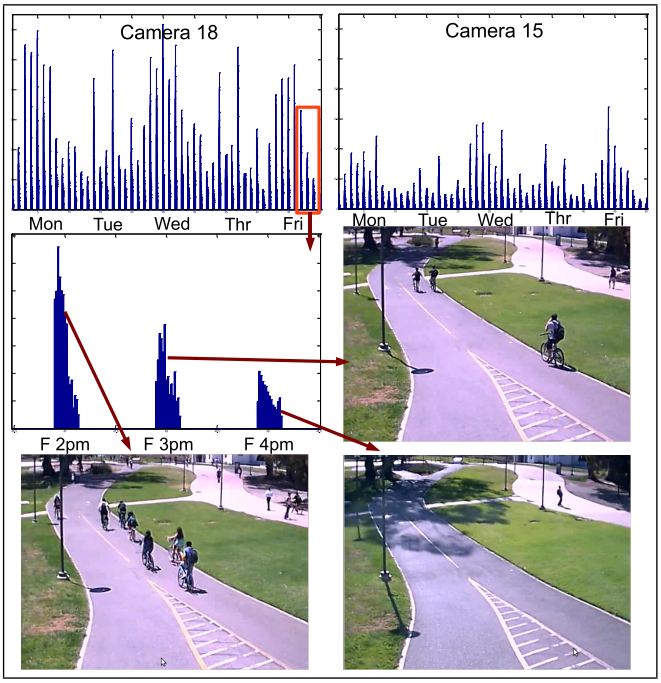

Figure 3: $As(T; c, h)$ with $T_w = 0.25$ and $c = 15, 18$

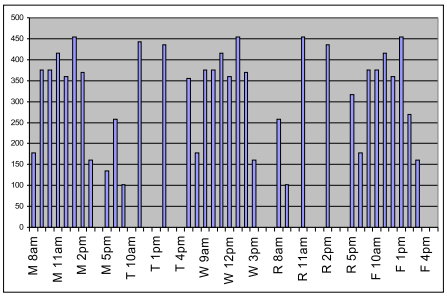

Figure 4: $N(b, h)$, b: building E

not easy to obtain given the large area being covered. Here we look at the problem from a different angle by utilizing building's schedule. We first start with a global/local camera classification step. If a camera observes the main campus entrance/exit, central bus station, or major bike paths/pedestrian walkways, we call it a global camera. Observation from these cameras should correlate well with the ensemble schedule of all buildings. To evaluate this, we use the canonical correlation between entire set of buildings $\mathbf{B}$ and camera c. Given the set of building $\mathbf{B}$, define random variable $X_{\mathbf{B}} = \sum_{b \in \mathbf{B}} N(b, h)$ with 50 observations ($h = 1 \ldots 50$) in a week.

For camera c, define random vector

$$Y_c = [As(T; c, h)]_{T=0,1,\ldots,\frac{20}{T_w}-1}, \quad (1)$$

with 50 observation ($h = 1 - 50$) in a week.

If activities observed by camera c are mainly caused by students going to buildings in $\mathbf{B}$, we expect a high canonical correlation $\rho(X_{\mathbf{B}}, Y_c)$. Canonical correlation analysis (CCA) finds two set of optimum basis vectors W_X and W_Y for $X_{\mathbf{B}}$and Y_c such that the correlation of the projections of them onto the basis vectors are maximized. CCA is defined

as,

$$\rho(X_{\mathbf{B}}, Y_c) = \frac{\text{Cov}(W_X X_{\mathbf{B}}, W_Y Y_c)}{\sqrt{\text{Cov}(W_X X_{\mathbf{B}}, W_X X_{\mathbf{B}})\text{Cov}(W_Y Y_c, W_Y Y_c)}}, \tag{2}$$

where Cov() means covariance. The solution to maximize $\rho(X_{\mathbf{B}}, Y_c)$ can be found with a set of eigenvalue equations [3]. The rational to use CCA is mainly due to different dimensions of $X_{\mathbf{B}}$ and Y_c. In addition, by projecting them into an optimal subspace, it minimizes the effect of pattern variations introduced by unknown information, and scale variation in the camera views etc. With a simple threshold th, we can then classify cameras into two categories, global and local cameras.

Figure 1.a shows classification result[2] with $th = 0.85$ and $T_w = 1$. Camera 10-12 and camera 15-18 cover the two main entrance/exit area (west and northwestern respectively); camera 13 covers the central bus stop; camera 19-23 cover the main bike path in the heart of the campus. All these cameras are correctly classified as global cameras, except for camera 20. We suspect the wrong classification of camera 20 is due to bad video quality (e.g. many short and missing recordings). All the rest of the cameras are classified correctly as local cameras as they only observe traffics linking to specific nearby buildings.

3.3 Fine-scale spatial-temporal estimation

For each classified local camera, we can further identify buildings which are strongly related to it. This can be done using similar techniques, i.e., computing CCA with X_b and Y_c. In particular, for each local camera c, we calculate its CCA $\rho(X_{\mathbf{b}}, Y_c)$ with each of its nearby buildings b. A simple threshold th_2 will give us a hint of what buildings are likely to cause activities seen by the camera c. Without exact knowledge (ground truth) on individual student's travel trajectory, it is hard to validate this method. Here we test our method on camera 26 to demonstrate the idea. With $th_2 = 0.8$ and $T_w = 1$, building B, D and E are found to be the ones related to camera 26. This corresponds very well with the finding we get after one week of human observation.

If the activities in camera c is due to students going to buildings in the set $\mathbf{B}_c$ (e.g., $\mathbf{B}_{26} = \{B, D, E\}$, we assume the following relation

$$As(T; h, c) = \sum_{b \in \mathbf{B}_c} \beta_b N(b, h) f(T; k_b, \theta_b), \tag{3}$$

where f defines movement patterns caused by student. $k_b > 0$ and $\theta_b > 0$ are their parameters. β_b weight the impacts of traffics to different buildings in the camera view. A gamma distribution, a typical probabilistic distribution for waiting time, is assumed to model f.

$$f(T; k_b, \theta_b) = T^{k_b - 1} \frac{e^{-T/\theta_b}}{\theta_b^{k_b} \Gamma(k_b)} (\text{for } 0 \le T < \frac{20}{T_w}) \tag{4}$$

Figure 5 shows the fitted $f(T; k_b, \theta_b)$ and estimated β_b for camera 26 with $T_w = 0.25$ (see Figure 3). Overall, it matches very well with our one week human observation. Some interesting findings are the followings: the peak traffics to each building occurs $6-7$ minutes ahead of class starting time; traffic flow toward building B and E are relatively continuous; traffic flow to building D appears to be a short impulse, which is because students tend to hang around in a corner coffee area (not observable by the camera) before entering class together. Other findings are omitted here due to the length constrain.

[2]Residential areas are located on the West and South of the UCSB campus. Major paths for non-motorized traffic are marked using red arrows.

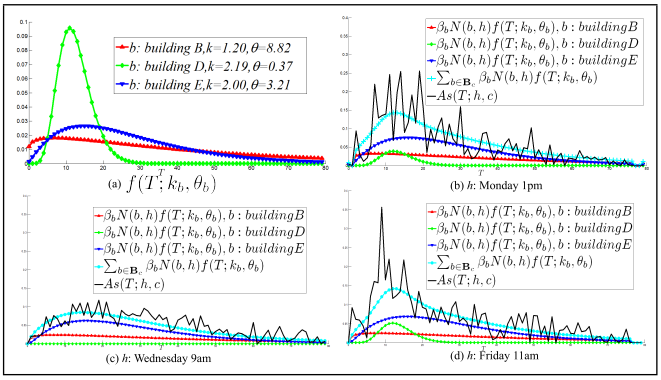

Figure 5: Fitted temporal pattern for camera $c = 26$.

4. CONCLUSIONS

In this paper, we have demonstrated the potential of spatial-temporal scene understanding through large camera network. In particular, we used simple vision algorithm to extract the crowdedness of the scene, and explored its relationship with campus activities based on class schedules. Similar analysis can be applied to other urban area with more diverse sources of information, e.g., theater schedule, mall/restaurant open time, office hours etc. In the future, we will deploy other sensors such as GPS to gain more quantitative ground-truths of human movement.

5. REFERENCES

[1] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004.

[2] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, 2008.

[3] C.Tofallis. Model building with multiple dependent variables and constraints. *Journal of the Royal Statistical Society Series D: The Statistician*, 48(3):1–8, 1999.

[4] J. Froehlich, J. Neumann, and N. Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *IJCAI'09*.

[5] M. C. Gonzalez, C. A. H. R., and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453:779–782, June 2008.

[6] D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari. What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *CVPR*, 2010.

[7] T. Kuo, Z. Ni, C. D. Leo, and B. Manjunath. Design and implementation of a wide area large-scale camera network. In *IEEE Workshop on Camera Networks*, 2010.

[8] L. Liu, A. Biderman, and C. Ratti. Urban mobility landscape: Real time monitoring of urban mobility patterns. In *CUPUM'09*.

[9] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, 2010.

[10] L. McNamara, C. Mascolo, and L. Capra. Media sharing based on colocation prediction in urban transport. In *MobiCom '08*.