ANOMALOUS ACTIVITY CLASSIFICATION IN THE DISTRIBUTED CAMERA NETWORK

Xiaotao Zou and *Bir Bhanu*
Center for Research in Intelligent Systems, University of California, Riverside, CA 92521

ABSTRACT

Unlike existing methods that used the human actions or trajectories to analyze the human activity in overlapping field-of-views, this paper proposes the appearance and travel time-based human activity classification in the camera network of non-overlapping field-of-views. The mixture of Gaussian-based appearance similarity model incorporates the appearance variance between different cameras to address changes in varying lighting conditions. To address the problem of limited labeled training data, we propose the use of semi-supervised Expectation-Maximization algorithm for activity classification. The human activities observed in a simulated camera network with nine cameras and twenty-five nodes are classified into one normal and three anomalous classes. A similar camera network is built and tested in real-life experiments, in which the proposed approach achieves satisfactory performance.

Index Terms— surveillance, activity analysis, camera network, semi-supervised learning

1. INTRODUCTION

Networks of video cameras are being envisioned for a variety of applications and many such systems are being installed. Thanks to the mass production of CCD or CMOS cameras and the increasing requirement in elderly assistance, security surveillance, traffic monitoring etc, a large number of video cameras has been deployed or are being constructed in our every-day life. However, most existing systems do little more than transmit the data to a central station where it is analyzed, usually with significant human intervention. As the number of cameras grows, it is becoming humanly impossible to analyze dozens of video feeds effectively. Therefore, we need methods that can automatically analyze the human activities in the video sequences collected by a network of cameras.

Suppose there are humans walking in the scene consisting of the conference room, the hallway, the patio and the doors to the stairs. Since the space is divided by the walls and rooms, the paths people can take are relatively constrained. Also, the travel times between the entry and exits of the key areas are relatively fixed depending on the characteristics of the pedestrians. The violation of the common paths and travel times constitutes the anomalous activities. For instance, there is someone taking the emergency exit of the conference room instead of the main door for convenience. It results in the unusual travel time between the conference room and the stairs much shorter than it is supposed to be. Another example is that someone climbs over the wall to circumvent the access control installed at the main entrance. Therefore, it seems as that the object suddenly appears at the door of the conference room without previously being detected at the main entrance. Other examples includes: suspicious long stay in the conference room and the sudden disappearance of the subject after showing at the entrance which means that the subject might hide somewhere. All these human activities mentioned above can be categorized to four main types: *break-in*, *stay*, *sudden appearance/disappearance* and *normal*. Among them, the first three anomalous activities require further attentions or human involvement.

There are many ways to classify the observed human activities and many sensor modalities available for this purpose such as imaging sensors [2][4], ultrasonic sensors [5], etc. Here we focus on anomalous human activity classification based on the widely used video cameras. One possible way is to track the objects (humans) across the overlapping field-of-views (FOVs) of different cameras and determine the types of human activities based on the observed tracks and travel times. However, the assumption of overlapping FOVs requires a huge number of cameras to cover a large area. The data volume increases exponentially along with the equipment cost making such an idea impractical. On the contrary, non-overlapping cameras overseeing the entry/exits in the environment greatly reduce the complexity of the surveillance system. However, the data correspondence problem also arises since there are multiple objects moving in the space and there exist "blind" areas or "gaps" between the FOVs.

We build upon these ideas to develop a framework for analyzing the activity patterns of a group of pedestrians given the inferred network topology and appearance similarity distribution. This paper uses the appearance similarity and travel times observed from much fewer cameras with non-overlapping FOVs to classify the human activities into four different classes: *normal*, *break-in*, *stay*, and *sudden appearance/disappearance*. This paper employs the color histogram-based appearance similarity to establish the correspondence between departure and arrivals at different nodes, and use the statistical model of appearance similarity to incorporate the uncertainty and variance of appearances between different FOVs under varied lighting.

Moreover, for a traditional learning-based classification scheme, sufficient labeled training data is the prerequisite of

satisfactory classification performance. However, it is really expensive to manually label a large volume of video sequences. Thus, we propose a semi-supervised Expectation-Maximization (SS-EM) algorithm to classify the human activities on the limited labeled data.

Main contributions of the paper are summarized as: (1) classification of normal and anomalous activities in non-overlapping FOVs using appearance similarity and travel time; (2) semi-supervised learning to identify anomalous activities in the activity behavior space; (3) mixture of Gaussian-based statistical model of appearance similarity for correspondence.

The paper is organized as follows: first, we discuss the color histogram-based appearance similarity and its statistical model based on Gaussian mixture model (GMM) in 2.1, followed by the introduction of the SS-EM algorithm in 2.2. Then, we present the anomalous human activity classification method by using the appearance similarity and travel time, and the SS-EM approach on the limited labeled training data. In Section 3, we show extensive simulation results for abnormal activity classification and the real-life experiment results are also presented. Finally, we conclude the paper in Section 4.

2. TECHNICAL APPROACH

The technical approach proposed in this paper is illustrated in Fig. 1. Like other learning-based classification methods, it has "training" and "testing" phases. In training, the network topology and appearance similarity distributions between different nodes are provided. The appearance similarity and travel time are first extracted from the training sequence. Due to the cost of manual labeling, only a limited number of labeled data and a large portion of unlabeled data are used in the SS-EM algorithm to estimate the GMM. The SS-EM algorithm is followed by a Bayesian classifier, and the classification results are evaluated with the ground truth which are fed back to the SS-EM to tune the parameters. The SS-EM is terminated if the fitness criterion is met. Otherwise, the iteration continues. In the testing, the estimated GMM is used by the Bayesian classifier to tell the anomalous from the normal activities.

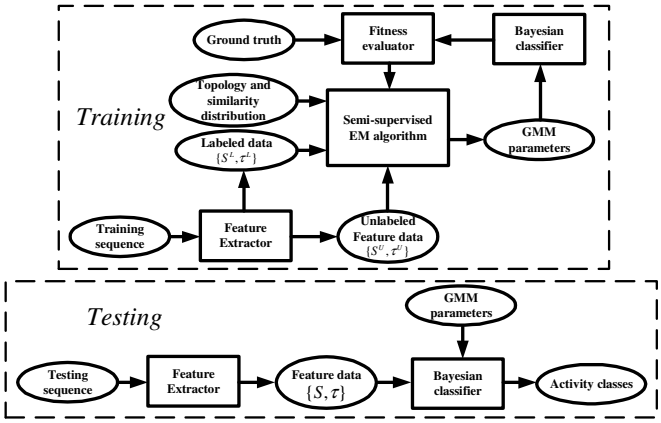


Figure 1. The dataflow diagram of the proposed SS EM-based human activity classification method.

2.1. Feature Extraction

Suppose the link from node i to node j is established in the network topology [3][6], and we are observing human activities of this specific link. By using existing motion detection and tracking techniques, we observe objects departing at node i and arriving at node j as temporal sequences $X_i(t)$ and $Y_j(t)$, respectively. The travel time variable τ is calculated as the difference between them.

$A_{X,i}(t)$ and $A_{Y,j}(t)$ are the observed appearances in the departure and arrival sequences, respectively. The color histograms in the hue and saturation space, i.e., h and s, respectively, are calculated on the normalized appearances $A_{X,i}(t)$ and $A_{Y,j}(t)$, respectively [6]. We use a k-component mixture of Gaussian distributions to model the color histogram similarity between the two appearances:

$$P_{app} = P(h_X - h_Y, s_X - s_Y \mid X \leftrightarrow Y) = \sum_{m=1}^{k} a_m \cdot N(h_X - h_Y, s_X - s_Y \mid \mu_m, \sigma_m^2) \quad (1)$$

where k is the number of components, α_m is the weight, μ_m and σ_m^2 are the mean and variance of the mth Gaussian component, and "$X \leftrightarrow Y$" means that they correspond to the same subject.

2.2. Semi-supervised EM Algorithm

Transductive learning combines both labeled and unlabeled data in training. In this scheme, labeled data provide the initialization and validation of the classifier, and the unlabeled data capture the statistical characteristics of the dataset and boosts the classifier. Various transductive learning methods have recently emerged. One approach dealing with labeled and unlabeled data from Gaussian mixture model is to modify the mixture log-likelihood function as the combination of two terms: one for labeled data and the other for unlabeled ones. Dong and Bhanu [1] present a short-term and long-term semi-supervised EM (SS-EM) based concept learning algorithm for content-based image retrieval. The assumption of Gaussian distribution provides good results for image retrieval. Our approach employs the SS-EM algorithm presented in [1].

First, let we briefly recall the procedures of the EM algorithm, in which the Expectation (E) and Maximization (M) steps are iterated until the termination.

- E-step: Compute the conditional expectation of the complete log-likelihood, given $\mathcal{X}$ and the current estimate $\hat{\theta}(t)$. We priorly know some binary component-indicator vectors Z such that: $z_{ji} = 1$, if $i = h$; or $z_{ji} = 0$, otherwise, where $i = 1, 2, \ldots, c$. h denotes the h^{th} component in the mixture of Gaussians. The result is the so-called Q-function:

$$Q(\theta, \hat{\theta}(t)) \equiv E\left[\log p(x, \mathcal{Y} \mid \theta) \mid x, \hat{\theta}(t)\right] = \log p(x, z \mid \theta) \quad (2)$$

In this equation, because of the linearity of $\log p(x,\mathcal{Y} \mid \theta)$ with respect to $\mathcal{Y}$, we only need to compute the conditional expectation $z \equiv E[\mathcal{Y} \mid x, \hat{\theta}(t)]$.

- *M*-step: Update the parameter estimates in the case of ML estimation according to:

$$\hat{\theta}(t+1) = \arg\max_{\theta} Q(\theta, \hat{\theta}(t)) \tag{3}$$

In SS-EM, the log-likelihood function is modified to incorporate both the labeled and unlabeled data:

$$\begin{aligned} &\log p(x, z \mid \theta) \\ &= \sum_{j=J^u} \sum_{i=1}^{c} z_{ji} \log\{\pi_i f_i(x_j \mid \theta_i)\} + \sum_{j \in J^L} \log\{\pi_h f_h(x_j \mid \theta_h)\} \end{aligned} \tag{4}$$

In the above equation, for the unlabeled data, the probability of an individual data point (the first term) is the sum of total probability over all classes. However, for the labeled data, the generating component h is already given and we just need to refer to the corresponding one (as in the second term) instead of all mixture components. The probabilistic indicator vector for the unlabeled data Z_{ji} (for $j \in J^u$) is the expected values according to the current parameter estimate $\tilde{\theta}(t)$.

Next we describe a weighted SS-EM method where the influence of the unlabeled data is modulated in order to control the extent to which EM performs unsupervised clustering. A new parameter λ is introduced into the log-likelihood function which balances the contributions of the unlabeled and labeled data to parameter estimation:

$$\begin{aligned} &\log p(x, z \mid \theta) \\ &= \lambda \cdot \sum_{j=J^u} \sum_{i=1}^{c} z_{ji} \log\{\pi_i f_i(x_j \mid \theta_i)\} + \sum_{j \in J^L} \log\{\pi_h f_h(x_j \mid \theta_h)\} \end{aligned} \tag{5}$$

Notice that when λ is close to zero, the unlabeled data will have little influence on the parameter estimation. When λ is close to one, each unlabeled data point will have almost the same influence as the labeled data, which is the same as the traditional EM algorithm.

2.3. Anomalous Activity Classification

In this section, we will employ the SS-EM algorithm combined with a naïve Bayes classifier for anomalous activity classification. When the naïve Bayes classifier is given just a small set of labeled training data, classification accuracy will suffer because variance in the parameter estimates of the generative model will be high. However, by augmenting this small labeled set with a large set of unlabeled data and combining these two with the framework of SS-EM, we can improve the parameter estimates and the classification accuracy significantly.

The object similarity distribution $P_{similarity}(S|X,Y)$, which has already been obtained in topology inference [6], can be combined with the transition time distribution $P(\tau)$ to identify the anomalous activity patterns. Under the assumption that the object similarity S and the transition time τ are independent, the joint distribution of object appearance similarity and transition time can be expressed as the product of $P_{similarity}(S)$ and $P(\tau)$:

$$P(S,\tau) = P_{similarity}(S \mid X(t), Y(t+\tau)) \cdot P(\tau) \sim \sum_k a_k \cdot N(\mu_k, \Sigma_k) \tag{6}$$

The anomaly activity patterns under study include *normal*, *break-in*, *stay* and *sudden appearance/ disappearance*, which are represented by the four components in the GMM. The class '*sudden appearance/disappearance*' means that either the subject never shows up at node *i* before appearing at node *j* or it never show up at node *j* after leaving node *i*. We do not distinguish between *sudden appearance* and *sudden disappearance* since both of them are of the same behavior type and distributed in the same area of the activity behavior space. The other 'suspicious' activities, such as *break-in* and *stay*, indicates that the subjects spend too much or too little time w.r.t. the expected transition time τ.

The GMM parameters are learned on the labeled and unlabeled data by using the SS-EM algorithm described before. Then, in the testing phase, the estimated Gaussian mixture model is used by the naïve Bayes classifier for anomalous activity classification.

3. EXPERIMENTAL RESULTS

3.1. Simulations

The simulation is based on *a priori* learned network topology shown in Fig. 2 (a). The simulated network has 18 departure/arrival nodes and 13 valid directed links. Some nodes, e.g., node 11, function as both 'departure' and 'arrival.' The traffic data of 100 points is generated by a *Poisson*(0.1) departure process, and the transition time follows Gamma distributions, e.g., *Gamma*(100, 5), *Gamma*(25, 2.5), etc. For simplicity, we only employ the appearance similarity and its probability P_{app} is modeled by a univariate Gaussian.

The simulated feature points in the activity behavior space are shown in Fig. 2 (b). We can find the four areas (marked as different colors) corresponding to the four different human activities. However, the distribution between these four areas is not balanced, i.e., most of the data points (~90%) are concentrated in the 'normal' region. It complies with the real-life environment. The first three classes are compact and their corresponding Gaussian components in GMM match them very well; however, the class *sudden appearance/disappearance* has a rough shape of Gaussian spanning over a large area because of the sparse data points in this class.

We randomly split the whole dataset for training and testing: 50% for training and 50% for testing. Among the training data, 30% are manually labeled and the other 70% remain unlabeled. The Gaussian mixture model estimated by the SS-EM algorithm is shown in Fig. 2(c). We can find that it capture the ground-truth GMM. The classification accuracy of the Bayes classifier is 99%. For comparison, we run an unsupervised k-means clustering algorithm on the same dataset. The number of clusters is initialized as four and the identified centers of the clusters are marked in Fig. 2 (d). We can find that they deviate significantly from the ground truth centers because of the unbalanced distribution

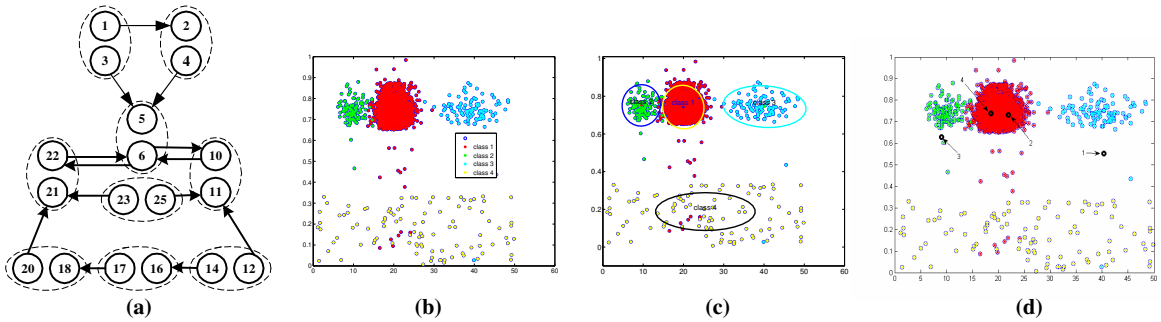


Figure 2. (a) The topology of the simulated camera network with 25 nodes and 13 directed links; (b) The simulated feature points in the activity behavior space; (c) the estimated Gaussian Mixture Model by the semi-supervised EM algorithm, which captures the property of the simulated data; and (d) the estimated class centers by using the unsupervised K-means clustering, which severely deviate from the true class centers.

of data points and the unsupervised learning style. Its classification accuracy is only about 27%.

3.2. Real-life Experiments

We construct a similar camera network overseeing one floor of a campus building, in which three are monitoring doors, one at the corner of the aisle and one covering part of the patio. The camera layout and network topology are shown in Fig. 3.

The observed video sequences with subjects arriving at difference nodes (i.e., FOVs of cameras) are displayed in Fig. 4. In our experiment, there are totally ten human subjects walking in the scene, and six times in-and-out of the FOVs for each subject. According to the footage, there are ten occurrences of anomaly: three *break-in*, three *stay* and four *sudden appearance/disappearance*. Our proposed approach has identified all these anomalous activities, i.e., the true alarm rate of 100%. Meanwhile, there are two false anomalous cases, shown in Fig. 4(b)(c), mainly due to the low resolution of the observations and the lack of strong biometric traits. The false alarm rate is 2/(60-10) = 4%.

4. CONCLUSIONS

This paper proposes to use the appearance similarity and travel times observed from much fewer cameras with non-overlapping FOVs to classify the human activities into four different classes: *normal*, *break-in*, *stay*, and *sudden appearance/disappearance*. To alleviate the problem of limited labeled training data, we propose to use the SS-EM algorithm for anomalous activity classification. The normal and abnormal activities observed in the simulated and real-life multi-camera network are tested.

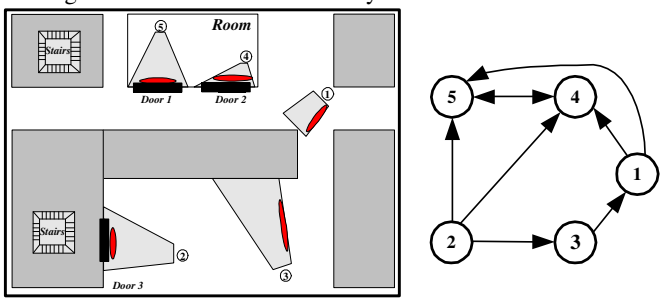


Figure 3. (Left) Experiment setup of the camera network showing their locations, FOVs (shade areas), and entry/exit points of the cameras (red ellipses), and (right) the topology of the real-life camera network.

ACKNOWLEDGEMENTS

This work was supported in part by NSF grants 0551741 and 0622176. The contents of the information do not reflect the position or policy of the US Government.

5. REFERENCES

[1] A. Dong and B. Bhanu, "Active Concept Learning in Image Databases," *IEEE Trans. SMC, Part B: Cybernetics*, 35(3): pp. 450- 465, June 2005.

[2] O. Jarved, Z. Rasheed, K. Shafique, and M. Shah, "Tracking Across Multiple Cameras with Disjoint Views," *Proc. ICCV*, pp. 1024-1029, June 2003.

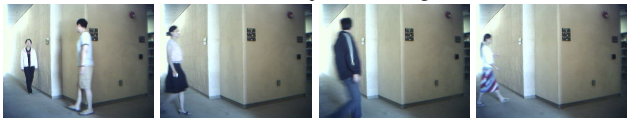
[3] D. Makris, T. Ellis, and J. Black, "Bridging the Gaps Between Cameras," *Proc. CVPR*, Vol. 2: pp. 205-210, 2004.

[4] D. Makris and T. Ellis, "Learning Semantic Scene Models from Observing Activity in Visual Surveillance," *IEEE Trans. SMC, Part B: Cybernetics*, 35(3): pp. 397-408, June 2005.

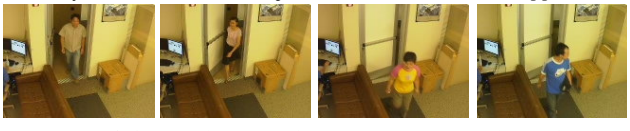
[5] C. Town, "Fusion of Visual and Ultrasonic Information for Environmental Modeling," *Proc. OTCBVS*, pp. 115-122, 2004.

[6] X. Zou, B. Bhanu, B. Song, and A. K. Roy-Chowdhury, "Determining Topology in a Distributed Camera Network," *Proc. ICIP*, pp. 911- 914, Sep. 2007.

(a) Camera 1: four subjects entering the FOV.



(b) Camera 3: four subjects entering the FOV. Note that the third from the left is falsely identified as other objects. It is a case of *sudden appearance*.



(c) Camera 5: four subjects entering the FOV. The one in the far right is falsely identified as another object. It is a case of *sudden appearance*.

Figure 4. The example frames captured by different cameras in the camera network with people entering the FOVs.