Visual Path Prediction in Complex Scenes with Crowded Moving Objects

YoungJoon Yoo[1], Kimin Yun[1], Sangdoo Yun[1], JongHee Hong[1], Hawook Jeong[2] and Jin Young Choi[1]
[1]Perception and Intelligence Lab., School of ECE, ASRI, Seoul National University
[2]Samsung Electronics Co., Ltd
[1]{i0you200, ykmwww, yunsd101, hundaeding, jychoi}@snu.ac.kr [2]hawook.jeong@samsung.com

Abstract

This paper proposes a novel path prediction algorithm for progressing one step further than the existing works focusing on single target path prediction. In this paper, we consider moving dynamics of co-occurring objects for path prediction in a scene that includes crowded moving objects. To solve this problem, we first suggest a two-layered probabilistic model to find major movement patterns and their co-occurrence tendency. By utilizing the unsupervised learning results from the model, we present an algorithm to find the future location of any target object. Through extensive qualitative/quantitative experiments, we show that our algorithm can find a plausible future path in complex scenes with a large number of moving objects.

1. Introduction

Scene understanding is an essential topic in the computer vision area, but lots of challenges remain in scene understanding research. In particular, the prediction of the future behavior of an object requires a highly intellectual inference regarding a scene structure and the objects' dynamics in a scene. The research on the visual prediction problem is in an initial stage because the problem requires a high level of inference on visual scenes, but the current scene understanding algorithms do not have the capability to infer a complex scene like a human. The current research is limited to specific problems such as occluded part prediction and future path prediction of an object in a scene, etc., as will be described in the related works in section 2. This paper aims to progress in the research on the future path prediction problem. Recently, a couple of works on the future path prediction problem have been presented [43, 21]. Kitani *et al*. [21] defined their specific prediction problem as finding the future trajectory of an object in an arbitrary location given the semantic structure of the scene. Walker *et al*. [43] proposed an algorithm to infer the shape and location changes of the representative patches considering the semantic structures of the scene after detecting the essential patches. The existing prediction works do not consider the reciprocal actions among moving objects in a scene.

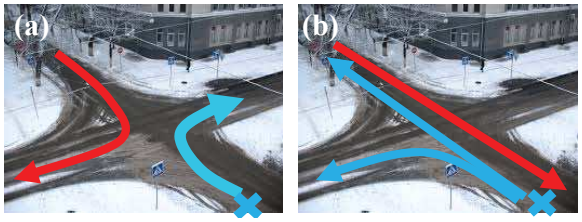

Figure 1. Cross-street which includes diverse movement patterns. To predict the future location of a target object at blue x point, finding the co-occurring movement pattern is required. The blue line shows the future trajectory of the object at blue x point and red line denotes the co-occurring movement patterns of other objects at the moment. If the turn right pattern (red) is dominant as in (a), the object of our interest will goes right and it will goes straight or left if straight pattern (red) occurs as in (b).

Even though their works pioneer the visual prediction field, the algorithms do not show satisfactory results in complex scenes, like cross-streets, where many objects interact with each other. In a cross-street, objects moves differently depending on the traffic light, even when they start from the same location in the cross-street.

Figure 1 is an example of this type of scene, which includes diverse movement patterns such as going straight and turning left. To predict the future location of a target object in this kind of scene, it is necessary to consider the movement patterns that occur at the prediction moment. For example, the object at the blue 'x' point will go to different destinations with respect to the co-occurring movement patterns (red line). If other objects (red) move right, as in (a), the object (blue) should go right in order to avoid collision. The target object (blue), likewise, will go straight or turn left when the co-occurring movement pattern (red) is straight as in (b). This example indicates that if we do not consider the dynamics of other co-occurring objects, we may get an inadequate predicted path, which may give rise to a collision.

In this paper, for one step of progress beyond the existing works, we propose a novel path prediction algorithm, which considers the moving dynamics of co-occurring objects. To the best of our knowledge, this is the first attempt in the path prediction research field. We develop a new unsupervised

Bayesian learning model that extracts typical movement patterns of objects and relationships from among the patterns to solve the prediction problem. The proposed model combines a topic mixture model [4] and the Gaussian mixture [30] hierarchically, which learns movement patterns as well as their interactions by utilizing the feature tracking results. However, the hierarchical combination of these two mixture models is not mathematically straightforward because the Guassian distribution is not a conjugate prior [8] of multinomial (topic) distribution, and so the posterior distribution of the combined model cannot be derived. Hence, this kind of combination has not been utilized despite its effectiveness. To resolve the problem, we introduce a mathematical trick to formulate a hierarchical topic-Gaussian mixture with satisfying the conjugate prior relation through an augmented variable. Then, we develop a deterministic path prediction algorithm utilizing the moving dynamics inferred by the proposed hierarchical topic-Gaussian mixture model. In this algorithm, we predict the future path of the target object by inferring the most plausible movement pattern for the object through analysis of the previous location of the object and moving dynamics of other co-occurring objects. We show that our algorithm finds a suitable future path of the target object through quantitative/qualitative experiments with widely used datasets [16, 5]. In particular, it is shown that, as expected, our method could predict the future path of objects and avoid collision with other co-occurring objects in the scene.

2. Related Works

Prediction capability is one of the essential indicators to measure the intellectual power and is extensively used to analyze intelligent behaviors of human [35] and animals [38]. Likewise, a variety of visual prediction research in computer vision has been conducted to measure the performance of scene understanding algorithms.

The existing visual prediction research includes many subcategories, such as predicting occluded parts [17], actions [42, 31, 22, 27], object dynamics [15, 1] and future path prediction [21, 43]. Recent path prediction research can be categorized into two approaches: path-planning-based approach and patch-appearance-based approach.

The first approach utilizes a path planning algorithm [25, 7, 28, 18, 26]. The approach uses statistical techniques such as inverse reinforcement learning [33, 46, 34] to find the optimal future path. Kitani *et al*. [21] first utilized the robot path planning algorithm to infer a point-wise future location of an object in a visual scene. The goal of this algorithm is to find a well-planned path for a target object with given scene structures such as roads, buildings, and so on. The object passes the appropriate area such as the pavement or road and avoids static obstacles in its way by following the induced path to reach the destination. To infer the predicted path, the algorithm first finds the cost for accessing each location in a scene and describes the cost via the reward map by utilizing the semantic segmentation result [32]. Then the algorithm extracts the optimal path which minimizes the overall cost by using inverse optimal control [33] and Markov decision process [2]. This approach is designed for single object movement prediction and does not consider possible collisions with other moving objects in a scene.

The second approach induces future changes of notable patches instead of locations of the target. In this case, inferring the representative patches is also a subproblem to be solved. Walker *et al*. [43] found the salient patches by applying recent mid-level patch-finding algorithms [11, 10, 14, 20, 37]. Then, they generated the weighted graph explaining the changes of the patches. The nodes of the graph represent the future locations and shapes of the patch. The weight is defined as a transition cost. The algorithm then finds the minimal weighted path by using Dijkstra's algorithm [9]. This path, starting from the initial node to termination, describes the changes of both shape and location. However, this algorithm has also been designed for single patches and does not reflect the dynamics of other moving objects in the scene.

Unlike the existing approaches, we propose a path prediction algorithm that reflects the movements of other co-occurring objects by using the novel hierarchical topic Gaussian mixture model. In the other research field, pattern analysis algorithms based on the probabilistic topic model [19, 41, 13, 16, 23, 24, 44, 40] also learn object dynamics in a scene to detect abnormal events. We highlight that our path prediction and motion pattern inference algorithms basically solve different problems. The existing topic model-based algorithms learns the regional patterns in a form appropriate to judge whether or not the target is moving in the typical regions. Meanwhile, the proposed HTGMM learns the object moving dynamics in the form of moving patterns together with their co-occurrences in a way that is adequate for future path prediction. Even the new model still provides the inference result in a quantized form because of the common limitations of the topic mixture [4, 6]. Therefore, we propose an efficient algorithm that induces the continuous future path from the quantized movement patterns and their relationship. To get the continuous path prediction, this paper transforms the quantized result into an energy potential map depicting the plausible paths in the form of valleys and predicts the future path by using the potential map. This prediction method is an essential contribution of the work together with the proposed HTGMM for the inference of moving object dynamics.

3. Proposed Method

The overall scheme of the proposed method is depicted in Figure 2. By analyzing the KLT trajectories [39], notable movement patterns are extracted from the scene by the proposed HTGMM. These patterns imply the semantic moving

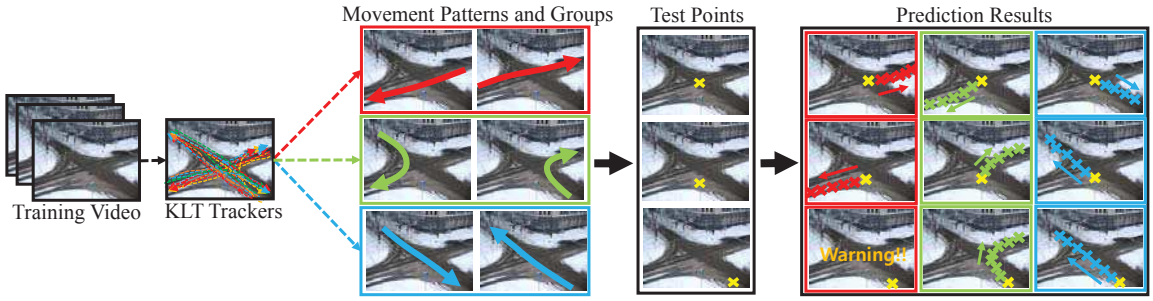

Figure 2. Overall scheme of the proposed method. The arrow in the scenes refers the movement pattern. Each pattern which occurs at the same time are located in colored boxes. Yellow x point is the location of the target object of which future path be predicted. Depending on the dominant group at prediction time, we induce different predicted paths.

dynamics of objects in a scene, such as going straight or turning right. Therefore, it is natural to expect that some patterns will occur at the same time according to their semantics. For example, we know that straight patterns going right and left in the separated lanes may usually occur simultaneously, as shown in the red lines in Figure 2. In this work, we divide the patterns into groups by considering the co-occurrence tendency among them. Each group, therefore, includes the patterns that may occur in the same time span. Utilizing this information, we predict the future trajectory of a target. As seen in Figure 2, depending on the dominant group at the prediction time, the predicted path can be different, even if the target starts from the same location. In the below sections, we give a detailed explanation of the proposed method.

3.1. Conversion of Input Trajectories

First, we convert KLT trajectories [39] into a set of words to be used as input features for the proposed probabilistic model. The sets of KLT trajectories are denoted by $T_l = \{(x_{lt}, y_{lt}) \mid t = 1, ..., N_T\}, l = 1, ..., N$. The term words, $w = \{w_i \mid i = 1, ..., N_w\}$, are defined as indices of the grids dividing a given scene. Then, each point (x_{lt}, y_{lt}) in a trajectory T_l is mapped to the word w_{lt} which indicates the grid including the point. N, N_T, and N_w respectively denote the total number of trajectories, the number of points in each trajectory, and the total number of the words w. Consequently, we can convert the trajectory T_l into the quantized form $T_l^{(w)} = \{w_{lt} \mid t = 1, ..., N_T\}$. In the below sections, we will write the quantized trajectory $T_l^{(w)}$ as T_l for convenience.

3.2. Hierarchical Topic-Gaussian Mixture Model

In this section, we introduce the unsupervised Hierarchical Topic-Gaussian Mixture Model (HTGMM). This model induces typical movement patterns and their co-occurrence types for a given quantized trajectory T_l. Figure 3 illustrates the proposed HTGMM in graphical representation. In a nutshell, the model learns K number of movement patterns into the topic mixture $\{\phi_k, q_k\}, k = 1, ..., K$, by utilizing the quantized KLT trajectories. Then, the patterns are clustered

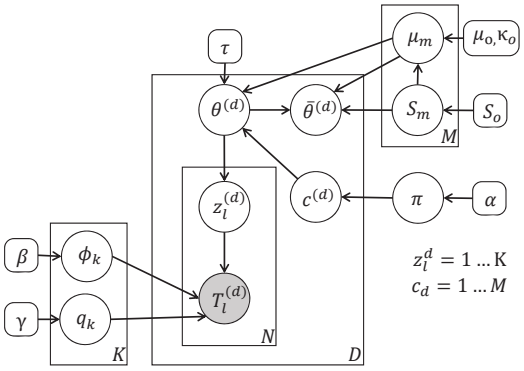

Figure 3. The proposed HTGMM. Each circle represents random variable. Empty circle denotes hidden variable and gray circle is an observed variable. Directed line represents conditional dependency between the circles and rectangle box means that the random variables and their dependency in the box are repeated with the number below the box.

into the mixture of M Gaussians, $\{\mu_m, S_m\}, m = 1, ..., M$, to infer M co-occurrence groups. The following gives the detailed description of the proposed HTGMM. First of all, to use overall quantized trajectories as input features to the model, we sort all the trajectories in order of ending times of the trajectories and evenly divide them into D number of chunks with N number of trajectories for each chunk. Through this procedure, the trajectories in a chunk occur in similar time span. The whole trajectories $T_l^{(d)}, d = 1, ..., D, l = 1, ..., N$, are used for the observed variables and clustered by the sets of random variables $\{\phi_k, q_k\}, k = 1, ..., K$, which indicates K number of patterns. $z_l^{(d)}$ is an indexing variable indicating the pattern type of the l-th trajectory in the d-th chunk, ranging from 1 to K. That is, it points out the pattern $\{\phi_k, q_k\}$ including $T_l^{(d)}$ among K patterns. ϕ_k is defined as the N_w dimensional random vector with multinomial distribution. The i-th element of ϕ_k indicates the probability that $k-$th pattern includes the i-th grid location, i.e., i-th word. ϕ_k learns the regional information of the k-th pattern. $q_k \in \mathbb{R}^{N_w \times N_w}$ denotes the word to word transition, i.e., direction, probability of k-th pattern. Consequently,

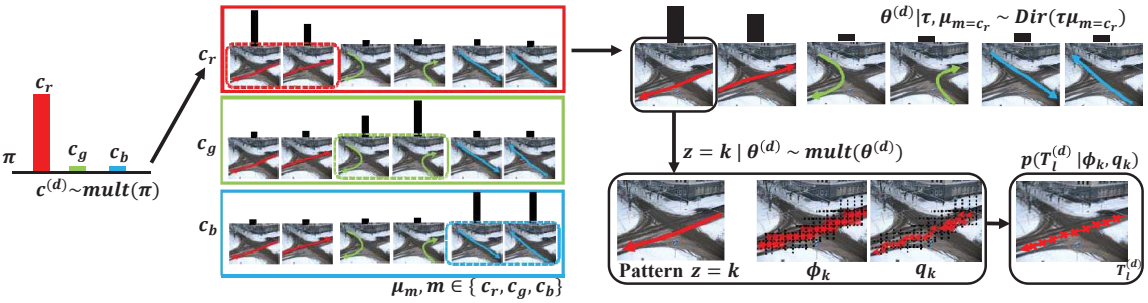

Figure 4. Explanation of proposed graphical model. The figure describes the generative procedure of the HTGMM.

given $z_l^{(d)} = k$, the probability that k-th pattern includes $T_l^{(d)} = \{w_{l1}^{(d)}, w_{l2}^{(d)}, ..., w_{lN_l^{(d)}}^{(d)}\}$, is given by

$$p(T_l^{(d)} \mid \phi_k, q_k) = \prod_{j=1}^{N_l^{(d)}} \phi_k(w_{lj}^{(d)}) \prod_{j=1}^{N_l^{(d)}-1} q_k(w_{lj}^{(d)}, w_{l(j+1)}^{(d)}). \tag{1}$$

Indexing variable $z_l^{(d)}$ is assigned by the multinomial distribution with parameter $\theta^{(d)} \in \mathbb{R}^K$ as

$$z_l^{(d)} \sim mult(\theta^{(d)}), \tag{2}$$

where $\sim$ means that the random variable $z_l^{(d)}$ has multinomial distribution with parameter of $\theta^{(d)}$, whereas $\theta^{(d)}$ represents the occurrence frequencies of the patterns in d-th chunk. In $\theta^{(d)}$, the entries with relatively high values give an information that the corresponding patterns have high tendency to occur simultaneously. It means that all $\theta^{(d)}, d = 1, ..., D$, give essential clues to find co-occurrence relationship of patterns. Therefore, we obtain M number of co-occurrence types by grouping $\theta^{(d)}$ into M clusters. To cluster the $\theta^{(d)}$, we set the mixture of M Gaussians $\{\mu_m, S_m\}$, $\mu_m \in \mathbb{R}^K, S_m \in \mathbb{R}^{K \times K}, m = 1, ..., M$. Accordingly, the entries of μ_m with high value represents major patterns in m-th group. The patterns in each group will occur at the same time with high probability. The example of obtained co-occurrence types is shown in Figure 4. $c^{(d)}$ is the indexing variable indicating one of Gaussian mixture, ranging from 1 to M. The indexing variable $c^{(d)}$ is assigned by multinomial distribution with parameter π as

$$c^{(d)} \sim mult(\pi). \tag{3}$$

However, since $\{\mu_m, S_m\}$ for the given $c^{(d)} = m$ is not a conjugate prior of $\theta^{(d)}$ [8], the posterior distribution of $\theta^{(d)}$ cannot be easily induced by using $\{\mu_m, S_m\}$ as a Gaussian prior of $\theta^{(d)}$. To resolve the difficulty, we additionally introduce an augmented variable $\bar{\theta}^{(d)} = f(\theta^{(d)})$ where $f(\cdot)$ is a deterministic mapping. It means that $\theta^{(d)}$ is converted to $\bar{\theta}^{(d)}$ with probability one. The performance depending on the choice of the mapping $f(\cdot)$ will be discussed in section 4. The Gaussian distribution can be the prior of $\bar{\theta}^{(d)}$ with any $f(\cdot)$ because $\bar{\theta}^{(d)}$ is not connected to $z_l^{(d)}$ as shown in Figure 3. After that, one of the Gaussian mixture selected by $c^{(d)} = m$ is defined as a prior of $\bar{\theta}^{(d)}$ i.e.,

$$\bar{\theta}^{(d)} \sim \mathcal{N}(\mu_m, S_m). \tag{4}$$

Note that $p(\bar{\theta}^{(d)} \mid \mu_m, S_m, \theta^{(d)}) = p(\bar{\theta}^{(d)} \mid \mu_m, S_m)$ given $p(\bar{\theta}^{(d)} | \theta^{(d)}) = 1$.

The procedure in the below is designed to let original $\theta^{(d)}$ assign $z_l^{(d)}$ indicating the dominant pattern in the group, $c^{(d)} = m$. To induce $\theta^{(d)}$ which reflects the co-occurring pattern information μ_m given $c^{(d)} = m$, τ and μ_m is defined as Dirichlet prior of $\theta^{(d)}$ i.e.,

$$\theta^{(d)} \sim Dir(\tau \mu_m). \tag{5}$$

Since Dirichlet prior $\tau \mu_m$ is pseudo count [3] of $\theta^{(d)}$, the entry of $\theta^{(d)}$ has higher value as the corresponding entry value of μ_m is larger. Furthermore, it is easy to induce the marginal distribution of $\theta^{(d)}$ because μ_m is conjugate prior of $\theta^{(d)}$. Hyper-parameters $\alpha, \beta, \gamma \in \mathbb{R}$ in Figure 3 are Dirichlet prior and $\mu_o \in \mathbb{R}^M, \kappa_o \in \mathbb{R}, S_o \in \mathbb{R}^{M \times M}$ are Nomral-Invert-Wishart prior [3] of Gaussian mixture $\{\mu_m, S_m\}$. These all parameters are conjugate priors of the corresponding random variables.

Joint pdf of the whole model is induced by combining the equations (1)-(5) altogether. However, it is impossible to get the exact posterior distribution of each variables because integral of the joint pdf is intractable due to the indexing variables z and c. Therefore, approximated inference methods are required to solve the problem. We use Gibbs sampling method [29] for inference of all the hidden variables in the proposed HTGMM. See the supplementary material for the detailed inference procedure.

3.3. Deterministic Method for Path Prediction

This section presents the path prediction method using the movement patterns and their co-occurrence groups learned by the proposed HTGMM. For this, we have to resolve two main problems. The first problem is that the movement patterns are described in quantized space. The other problem is that transition probability among words are defined only in the area of learned patterns. Therefore, we first suggest a method to expand the transition information of q_k into the entire word pairs. Then, we propose the final path prediction method inducing the future location x_{t+1} at time t in continuous domain given a previous target path $\mathbf{x}_t = \{x_1, x_2, ..., x_t\}$ in an iterative manner.

Relaxation of word to word transition: The word to word transition $q_k(w_i, w_j)$ indicates the direction of k-th move-

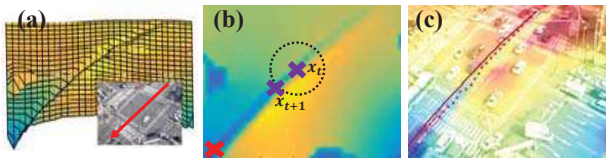

Figure 5. The result of expanded word to word transition. We obtain energy potential map in (a) by expanding word to word transition of the pattern in bottom of (a). The potential goes down from yellow to blue. We induce the potential map in continuous domain by bi-lnear interpolation in (b). Therefore, the sink point of the map (red 'x') indicates the destination of the pattern. The points at purple 'x' represent x_{t+1} and x_t. The figure (c) shows the example path prediction result.

ment pattern from i-th grid to j-th grid. The (w_i, w_j) is a word pair in a scene where the condition $q_k(w_i, w_j) \neq 0$ is satisfied. Since we do not have the transition information for all the word pairs, the total number of trained word pairs (w_i, w_j) is less then whole possible number of word pairs ${N_w}^2$. To expand the word to word transitions to whole word pairs, we employ an energy potential vector $\mathbf{y} = [y_1, y_2, ..., y_{N_w}]^T$. The y_i, y_j are defined so that $y_i - y_j = q_k(w_i, w_j)$. If we know the transition probabilities for R pairs of words, we can set R equations for each (y_i, y_j). The set of the equations can be expressed by sparse matrix form $A\mathbf{y} = b, A \in \mathbb{R}^{R \times N_w}, b \in \mathbb{R}^R$ which holds $A[r,i] = 1$,$A[r,j] = -1$ and $b[r] = q_k(w_i, w_j)$. $A[r,i]$ and $A[r,j]$ are (r,i) and (r,j) element of matrix A. Also, $b[r]$ is the r-th element of vector b. In most cases, A is not a full rank matrix. Accordingly, we can find a solution as $\mathbf{y} = (A^T A)^{-1} A^T \mathbf{b}$ by using pseudo inverse. Using the $\mathbf{y}$, we induce transition probabilities of whole word pairs in a scene. Figure 5 is an example illustrating the $\mathbf{y}$. The difference between y_i and y_j at each location denotes the possibility that the target moves from high potential position w_i to low potential position w_j. The Figure 5 shows that the potential value decreases as the target moves to the future locations.

Path Prediction in Continuous Domain: After finding the potential map $\mathbf{y}$, we iteratively update $\mathbf{x}_t$. The overall path prediction procedure has three steps. In the first step, we find the movement patterns adequate to the target object by using the inference results from HTGMM. The second step is the updating procedure for $\mathbf{y}_t$. In this step, we modify $\mathbf{y}$ to reflect the past trajectory of the target, $\mathbf{x}_t$. We denote $\mathbf{y}$ at time t as $\mathbf{y}_t$. In the last step, we estimate future location x_{t+1} of the target using the updated $\mathbf{y}_{t+1}$.

(1) Pattern selection step: This step begins with converting the past trajectory $\mathbf{x}_t = \{x_i \mid i = 1, ..., t\}$ into a quantized form $\mathbf{x}_t^{(q)} = \{w_i \mid i = 1, ..., t\}$ where w_i is a word including x_i. Then we select the pattern including $x_t^{(q)}$ according to the probability of selecting k-th pattern $\{\phi_k, q_k\}$ given the dominant pattern group c by employing the results from HTGMM as

$$p(\{\phi_k, q_k\} \mid \mathbf{x}_t^{(q)}, \mu_c) \propto p(\mathbf{x}_t^{(q)} \mid \{\phi_k, q_k\}) p(z = k \mid \mu_c). \tag{6}$$

The first term in the right-hand side of the equation can be obtained from the equation (1). It represents the probability that k-th movement pattern includes the target trajectory $\mathbf{x}_t$. The second term is a Dirichlet multinomial distribution over μ_c. The distribution is induced by marginalizing θ of $p(z = k \mid \theta) p(\theta \mid \mu_c)$ where $p(z = k \mid \theta)$ and $p(\theta \mid \mu_c)$ can be obtained from the equations (2) and (5). It is a tractable calculation because μ_c is a conjugate prior for θ. The second term leads to the selection of z indicating the frequently occurring pattern in the group c. The group c is determined by the maximum value of the posterior probability for μ_c in the HTGMM with given co-occurring KLT trajectories.

(2). Energy potential map update step: After selecting the pattern k, we update $\mathbf{y}_{t+1}^k$ using $\mathbf{y}_t^k$ and $\mathbf{x}_t$. We denote $\mathbf{y}_t^k$ as the potential vector $\mathbf{y}$ for k-th pattern at time t. To estimate $\mathbf{y}_{t+1}^k$ reflecting the trace $\mathbf{x}_t$, we first define $t-1$ terms in equation (7) from the $\mathbf{x}_t^{(q)}$, where y_i is the energy potential assigned for the word w_i in $\mathbf{x}_t$.

$$y_{w_{i+1}} - y_{w_i} = p, i = 1, ..., t-1. \tag{7}$$

Then, we add them into the rows of the matrix A, b used previously for calculating the potential vector. By solving the linear equation $A\mathbf{y} = b$ with modified A and b, we obtain a new vector $\mathbf{y}_c$ containing the future dynamics estimated from the past movements. We set p as a mean value of all $q_k(w_u, w_v) \geq 0$ in a scene. The vector $\mathbf{y}_{t+1}^k$ is updated by reflecting the $\mathbf{y}_c$ to the present state as

$$\mathbf{y}_{t+1}^k = (1 - \alpha)\mathbf{y}_t^k + \alpha \mathbf{y}_c, \tag{8}$$

where term α is a design parameter.

(3) Path prediction step: Now we finally find $\mathbf{x}_{t+1}$ using the $\mathbf{y}_{t+1}^k$. As seen in Figure 5, the map $\mathbf{y}_{t+1}^k$ forms a valley-like shape going down to the destination of the pattern k. Therefore, we find x_{t+1} by following the slope of the valley. To find new x_{t+1} in a continuous domain, we expand $\mathbf{y}_{t+1}^k$ into continuous space using bi-linear mapping [36]. $\mathbf{F}_{t+1}^k$ refers the continuous energy potential map obtained from $\mathbf{y}_{t+1}^k$. Then, we find the sink point x_s of the $\mathbf{F}_{t+1}^k$ which indicates the destination of the pattern k. The optimization formulation to find x_{t+1} is given by

$$\begin{aligned} x_{t+1} = & min_x \mathbf{F}_{t+1}^k(x), \\ & s.t. \ \|x - x_t\|_2 = \|x_t - x_{t-1}\|_2, \\ & \|x - x_s\|_2 \leq \|x_t - x_s\|_2. \end{aligned} \tag{9}$$

To find the minimal point in (9), we only need to navigate the points x lying in the circle $C(R, \theta)$ which $R = \|x_t - x_{t-1}\|_2$ and $-\pi \leq \theta \leq \pi$ with center x_t. To find x with minimal $F_{t+1}^k(x)$, we find inflection points by calculating θ satisfying the gradient $\nabla_\theta F_{t+1}^k(C(R, \theta)) = 0$ and choose the point with the minimum field value as the future location x_{t+1}. By increasing the time index t, we predict the future

location of target recursively and we terminate the recursive iteration when the distance between predicted point x_{t+1} and x_s is smaller than $\|x_t - x_{t-1}\|_2$ or x_{t+1} goes over the boundary of the scene.

4. Experimental Results

To validate the proposed algorithm, we compared the performance against the existing path prediction algorithms [43, 21]. Through the comparison, we have confirmed that the existing path prediction algorithms [43, 21] are not adequate for the crowded scenes which have a temporal pattern co-occurrence tendency. Also, to check our method's applicability to pedestrian moving patterns, we compared it with Yi's method [45]. In addition, to check the effects of the components of the proposed HTGMM model, we conducted extensive experiments to evaluate our algorithm by self-comparing its performance with that of three baseline algorithms designed by naive combinations of the existing topic and Gaussian models.

4.1. Dataset

For the experiments, we first used QMUL[16, 5], including cross road scenes and our own complex intersection (CI) dataset captured in a wide-intersection. These scenes include diverse moving object patterns and co-occurrence types governed by traffic signals. Furthermore, these scenes are very crowded, and it is hard to utilize semantic scene segmentation information as in the previous works [21, 43]. In addition, for the pedestrian data set, we adopted PWPD [45] which does not have explicit temporal groups among the movement patterns. The dataset captures a crowded indoor plaza scene in a subway station, and the movement of the objects is far less ordered compared to the QMUL, CI datasets.

4.2. Comparison Methods

First, we compared the prediction performance with two major existing path prediction algorithms [43, 21] for the QMUL, CI datasets. Walker's method [43] learns the transition probability among representative mid-level patches and predicts the shape and future position of the patch. Kitani's method [43] trains the reward function for each location given semantic segmentation results and finds the predictive path which minimizes the cost. For comparison, we measured the error between ground truth trajectory and predicted trajectories of each algorithms using modified Hausdorff distance (denoted by MHD in tables) [12] and Euclidean distance (denoted by ECD in tables). Since Walker's method automatically determines the patches for prediction, we generated ground truth trajectories for the selected patches. For the PWPD dataset, we compared the performance with Yi's method[45] which marks the state-of-the-art performance to the PWPD dataset. This method does not explicitly focus on predicting trajectories but can predict the possible destination region of objects in the scene by seeing half of the entire paths.

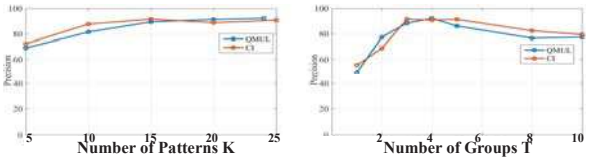

Figure 6. Precision Graph with respect to number of patterns K and number of groups M.

Second, in addition to the existing algorithms [43, 21, 45], we employed our own three baseline algorithms. In the first baseline algorithm, utilizing the movement patterns $\{\phi_k, q_k\}$ and $\theta^{(d)}$ obtained by the HTGMM, we simply inferred the co-occurrence of the movement patterns by clustering $\theta^{(d)}$ with a Gaussian mixture model. This baseline algorithm refers to 'B(1)'. The method naively breaks the proposed HTGMM into two independent models and infers the hidden variables in a greedy manner. The second baseline algorithm is designed with the same concept as the first baseline algorithm except for using $\bar{\theta}^{(d)}$ instead of $\theta^{(d)}$. The purpose of the second baseline is to show that only the simple mapping $\bar{\theta}^{(d)} = f(\theta^{(d)})$ does not give significant improvement of performance without the prior design as in the proposed HTGMM. The second baseline algorithm refers to 'B(2)'. The other baseline algorithm 'B(3)' assumes just one group. This means that the third baseline algorithm does not consider co-occurrence information. In addition, we added the prediction result obtained by humans to evaluate the prediction performance relative to human ability. Five human participants saw the training video three times repeatedly to learn the movement dynamics. They then predicted the future path from the same points given in the experiments for the proposed algorithm.

4.3. Qualitative Evaluation

To evaluate the robustness of design parameters, we tested our work with different parameters, namely the number of patterns K and the number of groups M. As seen in the left graph in Figure 6, our method is robust in relation to K unless the number is too small. M is a more sensitive parameter than K. In the traffic scenario, we observed that selecting three to five groups achieves the best performance. It is noticeable that the performance gap is less severe if we choose a value larger than the fitted parameters. Figure 7 shows the patterns and their co-occurrence types extracted by the proposed algorithm. Each pattern is illustrated by utilizing regional probability ϕ_k and the potential energy map $F^{(k)}$. The co-occurrence groups of the patterns are illustrated in the right four images in each row of Figure 7. By utilizing the results, we measured the pattern-trajectory matching accuracy, indicating whether a trajectory is matched to an appropriate pattern in the situation at the prediction time.

CI Dataset: We set the number of patterns, K, and the number of groups, M, to 15 patterns and three co-

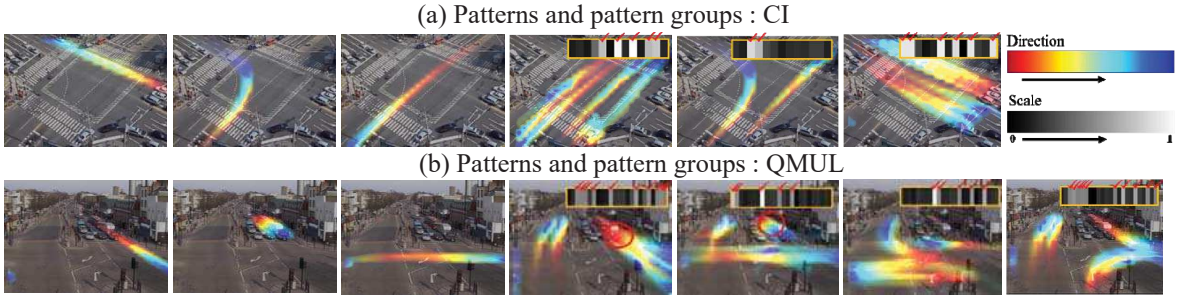

Figure 7. The inferred movement patterns and their co-occurrence groups. In each rows, three images in the left indicates the examples of movement patterns. Other images in the rightside depict their groups. The color of each pattern indicates the direction of the pattern, from red to blue. The bar in each picture in the rightside of each rows represents the μ_m of each Gaussian group. The gray-scaled color in the bar indicates the occurrence probability of a pattern, where a white color shows a high probability. It means that the white entries of the bar show the major patterns of the group in the corresponding picture. Best viewed in color.

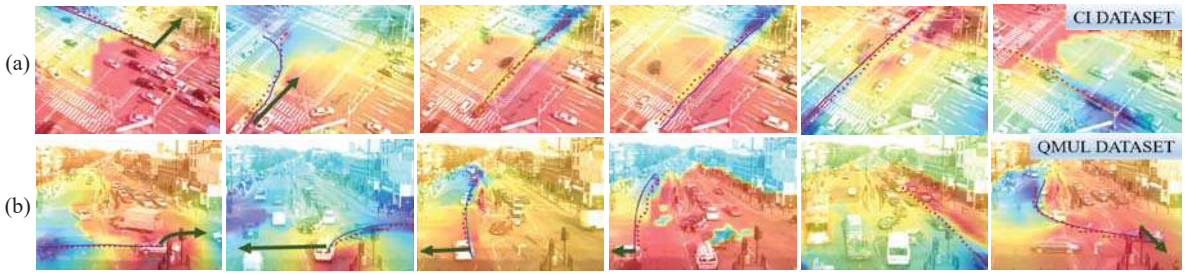

Figure 8. Illustration of diverse path prediction results in different groups. The solid lines indicate the ground truth trajectories and dot lines denote the predicted paths. The green arrows indicate the other possible directions if the co-occurrence groups are changed.

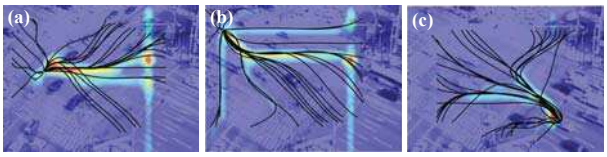

Figure 9. Path prediction results of Walkers' [43] for CI dataset.

occurrence groups, respectively, to learn the CI dataset. As shown in the right three images of Figure 7-(a), the proposed model can successfully make three groups with co-occurring patterns depending on the major co-occurrence types generated by traffic signals: horizontal straight, turning left with vertical straight, and vertical straight. By utilizing those patterns and groups, we conducted a prediction task and evaluated the prediction performance with 189 ground truth trajectories. As illustrated in Figure 8-(a), we can see that the predicted trajectories do not go toward moving objects (green arrow direction) considering co-occurrence group and arrive at the destination by following the valley obtained by the energy potential map and are matched to the ground truth. Conversely, as in Figure 9, the predicted path by [43] for CI data set guides cars to avoid other cars, which results in an erroneous prediction in crowded traffic conditions.

QMUL Dataset: For this dataset, we set K and M to 24 patterns and four co-occurrence groups, respectively, because the scene structure is more complicated. Figure 7-(b) represents the patterns and co-occurrence groups, extracted from the QMUL dataset. In Figure 7-(b), it is worth highlighting that the vertical straight patterns depicted by red circles in the first two groups are included in different co-occurrence groups even though they are passing the same region. Hence, their future paths will be different from each other depending on the movements of other objects. In other words, the object in the first pattern will keep going according to the vertical straight pattern, but the object in the second pattern will stop near the crosswalk region to avoid a collision with the horizontal movements. Figure 8-(b) shows the prediction results given the groups. We executed the prediction experiment and evaluated the performance with 246 ground truth trajectories. In this scene, there are many locations too complicated for choosing the pattern, but our algorithm successfully selects adequate patterns for prediction. For example, the trajectory in the first image and int the second image in Figure 8-(b) start from almost the same location, but the predicted path is completely different depending on the co-occurrence group that is dominant at the prediction time.

PWPD Dataset: We tested our method in the pedestrian walking path dataset [45] which captures complex dynamic crowd movements. The experimental results in the PWPD dataset [45] shows that the applicability of the proposed algorithm is not restricted to cross-road traffic scenes, but can be used for a more disordered situation. Since this scene does not include the temporal group, such as traffic controlled by traffic signals, we set the number of group M

Video	measure	Human	**Proposed**	**Proposed(2)**	W14(1)	W14(2)	K12	B(1)	B(2)	B(3)
QMUL	Precision	99.37	**92.14%**	**-**	-	-	-	67.36%	73.14%	49.58%
	MHD [12]	23.32	**23.38**	**11.65**	49.36	76.20	86.73	41.90	35.34	65.70
	ECD	45.71	**50.19**	**36.80**	85.5	115.29	107.43	71.47	59.51	88.05
CI	Precision	99.20	**91.49%**	**-**	-	-	-	63.29%	65.42%	55.31%
	MHD	21.22	**27.89**	**14.72**	62.03	115.51	127.62	45.04	43.91	49.68
	ECD	40.15	**44.95**	**28.60**	92.50	150.43	143.60	63.29	56.15	68.59

Table 1. Quantitative results of cross-street dataset. MHD indicates modified Hausdorff distance [12] and ECD denotes Euclidean distance. W14 refers to Walker's method[43] and K12 refers to Kitani's method[21]. B1,B2 and B3 indicates Baseline algorithms in section 4.2. W14(1) is mean value of the top 10% lowest error. W14(2) represents error of the path which has the highest probability. The result K12 is from the same configuration as W14(2).

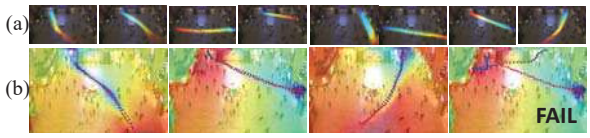

Figure 10. Qualitative Prediction Results for PWPD dataset [45]. (a) Extracted patterns of ours, (b) Our prediction results.

to 1 and the number of patterns K to 40 via experiments which were not sensitive. We used the object trajectories given by the author [45]. As shown in Figure 10-(a), our model successfully learned movement patterns. Figure 10-(b) describes examples of path prediction results. The results show that our model successfully predicts the future when the object(human) does not loiter, as in Figure 10-(b).

4.4. Quantitative Evaluation

First, we conducted a quantitative comparison of the algorithms[21, 43] with the videos proposed in the paper. Table 1 shows the comparison results. For Walker's method[43], we used the mid-level features trained by the car chase dataset and the CI, QMUL datasets. For Kitani's work[21], manual ground truth segmentation results were adopted. Although the conditions of the experiment were advantageous to them, our method yielded superior performance because the two methods are designed to avoid obstacles such as cars and lawns.

We also measured the performance of the algorithm and compared the results with the baseline methods as well as with human prediction. As shown in Table 1, the proposed algorithm outperformed the other baseline algorithms in both datasets. The result implies that the proposed method has a meaningful contribution compared to the naive use of the existing topic and Gaussian mixture models. The baseline algorithm 'B(1)' achieved better performance than the baseline algorithm 'B(3)' which does not group the patterns. However, since the group information learned by the first baseline algorithm was inaccurate, the performance improvement by GMM was insufficient. Considering that the baseline algorithm uses the same $\theta^{(d)}$ learned by the proposed HTGMM, we conclude that the performance jump of the proposed method in comparison with the baseline algorithm 'B(1)' validates our model's superior ability to group co-occurring patterns. The result of 'B(2)' implies that utilizing sigmoid function without the proposed conjugate prior design in HTGMM does not yield a good performance. Furthermore, the prediction result (MHD, ECD) from humans and 'Proposed(2)' shows that our algorithm has a comparable prediction ability to that of humans in view of distance error. The result of 'Human' in Table 1 is the average value for the five humans. Interestingly, even humans were confused in predicting the path of the target, which can go in multiple directions depending on the co-occurrence dynamics.

Video	measure	Y15(1)	Y15(2)	Y15(3)	Ours
PWPD	Precision	48%	38%	33%	**43.2%**

Table 2. Pedestrian destination results. Y15(1)[45] is the result which uses the stationary crowd factor. Y15(2),(3) are the baselines which do not, or naively use the factor.

Also, as shown in Table 2, our method achieved destination predicting performance comparable to the newest method [45] without employing the stationary crowd information, claimed to be the essential feature by Yi *et al*. [45] for analyzing a crowded scene like the PWPD dataset. It is noted that our method outperforms the other baselines of [45]: Y15(2), Y15(3), which do not utilize that factor.

5. Conclusion

In this paper, we have proposed a novel path prediction algorithm that considers the moving dynamics of co-occurring objects. To solve the problem, we first designed two-layered probabilistic model to extract the major movement patterns and their co-occurrence groups in a scene. Utilizing the result from the proposed model, we have presented an effective path prediction method. By extensive qualitative/quantitative experiments, we have shown that our algorithm can predict the future paths of objects in complex scenes including many moving objects and changing situations such as cross streets with traffic lights. This paper explores a meaningful progress in path prediction research in that the proposed algorithm considers the other co-occurring objects as well as the target itself.

6. Acknowledgment

This work was partly supported by the ICT RD program of MSIP/IITP[B0101-15-0552, Development of Predictive Visual Intelligence Technology] and the Brain Korea 21 Plus Project.

References

[1] I. Ardiyanto and J. Miura. Human motion prediction considering environmental context. In *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on*, pages 390–393. IEEE, 2015.

[2] R. Bellman. A markovian decision process. Technical report, DTIC Document, 1957.

[3] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[4] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[5] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang. Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2909–2917, 2015.

[6] A. Daud, J. Li, L. Zhou, and F. Muhammad. Knowledge discovery through directed probabilistic topic models: a survey. *Frontiers of computer science in China*, 4(2):280–301, 2010.

[7] D. Devaurs, T. Siméon, and J. Cortés. Efficient sampling-based approaches to optimal path planning in complex cost spaces. In *Algorithmic Foundations of Robotics XI*, pages 143–159. Springer, 2015.

[8] P. Diaconis, D. Ylvisaker, et al. Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281, 1979.

[9] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[10] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *Advances in Neural Information Processing Systems*, pages 494–502, 2013.

[11] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012.

[12] M.-P. Dubuisson and A. K. Jain. A modified hausdorff distance for object matching. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 566–568. IEEE, 1994.

[13] R. Emonet, J. Varadarajan, and J.-M. Odobez. Extracting and locating temporal motifs in video scenes using a hierarchical non parametric bayesian model. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3233–3240. IEEE, 2011.

[14] I. Endres, K. Shih, J. Jiaa, and D. Hoiem. Learning collections of part models for object recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 939–946. IEEE, 2013.

[15] D. F. Fouhey and C. L. Zitnick. Predicting object dynamics in scenes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2027–2034. IEEE, 2014.

[16] T. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1165–1172. IEEE, 2009.

[17] D.-A. Huang and K. M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *Computer Vision–ECCV 2014*, pages 489–504. Springer, 2014.

[18] M. Jalalmaab, B. Fidan, S. Jeon, and P. Falcone. Model predictive path planning with time-varying safety constraints for highway autonomous driving. In *Advanced Robotics (ICAR), 2015 International Conference on*, pages 213–217. IEEE, 2015.

[19] H. Jeong, Y. Yoo, K. M. Yi, and J. Y. Choi. Two-stage online inference model for traffic pattern analysis and anomaly detection. *Machine vision and applications*, 25(6):1501–1517, 2014.

[20] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 923–930. IEEE, 2013.

[21] K. Kitani, B. Ziebart, J. Bagnell, and M. Hebert. Activity forecasting. *Computer Vision–ECCV 2012*, pages 201–214, 2012.

[22] H. Koppula and A. Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 792–800, 2013.

[23] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1446–1453. IEEE, 2009.

[24] D. Kuettel, M. D. Breitenstein, L. Van Gool, and V. Ferrari. What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1951–1958. IEEE, 2010.

[25] S. M. LaValle. *Planning algorithms*. Cambridge university press, 2006.

[26] S. M. LaValle and J. J. Kuffner. Randomized kinodynamic planning. *The International Journal of Robotics Research*, 20(5):378–400, 2001.

[27] K. Li and Y. Fu. Prediction of human activity by discovering temporal sequence patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(8):1644–1657, 2014.

[28] Z. W. Lim, D. Hsu, and W. S. Lee. Adaptive informative path planning in metric spaces. In *Algorithmic Foundations of Robotics XI*, pages 283–300. Springer, 2015.

[29] J. S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

[30] J.-M. Marin, K. Mengersen, and C. P. Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, 25(16):459–507, 2005.

[31] B. Minor, J. R. Doppa, and D. J. Cook. Data-driven activity prediction: Algorithms, evaluation methodology, and applications. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 805–814. ACM, 2015.

[32] D. Munoz, J. A. Bagnell, and M. Hebert. Stacked hierarchical labeling. In *Computer Vision–ECCV 2010*, pages 57–70. Springer, 2010.

[33] A. Y. Ng, S. J. Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, pages 663–670, 2000.

[34] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. *Urbana*, 51:61801, 2007.

[35] D. L. Schacter, D. R. Addis, and R. L. Buckner. Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*, 8(9):657–661, 2007.

[36] R. J. Schalkoff. *Digital image processing and computer vision*, volume 286. Wiley New York, 1989.

[37] S. Singh, A. Gupta, and A. Efros. Unsupervised discovery of mid-level discriminative patches. *Computer Vision–ECCV 2012*, pages 73–86, 2012.

[38] E. L. Thorndike. Animal intelligence: an experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, 2(4):i, 1898.

[39] C. Tomasi and T. Kanade. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

[40] J. Varadarajan, R. Emonet, and J.-M. Odobez. Bridging the past, present and future: Modeling scene activities from event relationships and global rules. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2096–2103. IEEE, 2012.

[41] J. Varadarajan, R. Emonet, and J.-M. Odobez. A sequential topic model for mining recurrent activities from long term video logs. *International journal of computer vision*, 103(1):100–126, 2013.

[42] T.-H. Vu, C. Olsson, I. Laptev, A. Oliva, and J. Sivic. Predicting actions from static scenes. In *Computer Vision–ECCV 2014*, pages 421–436. Springer, 2014.

[43] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3302–3309. IEEE, 2014.

[44] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(3):539–555, 2009.

[45] S. Yi, H. Li, and X. Wang. Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3488–3496, 2015.

[46] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, pages 1433–1438, 2008.