EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE, pp. 776-781, VIENNA, AUSTRIA, AUGUST 1992

On the Visual Expectations of Moving Objects

Shaogang Gong* **Hilary Buxton***
Computer Science Department, Queen Mary and Westfield College,
Mile End Road, London E1 4NS, England

Abstract. Meaningful objects in a scene move with purpose. In this work, a Hidden Markov Model is used to represent the "hidden" regularities behind the apparent object movement in a scene and reproduce such regularities as "blind" expectations. By adapting the weights on beliefs with new visual evidence, an Augmented Hidden Markov Model is used to produce dynamic expectations of moving objects in the scene. Keywords: Visual observation, tracking, visual expectation and visual attention, learning, probabilistic belief network, augmented Hidden Markov Model.

1 Introduction

Vision is highly selective [4, 1], purposive [5] and active [3]. Task knowledge and the nature of the scene often define the visual attention and allow us to ignore the irrelevant [4]. In visual observation, understanding and interpreting moving objects in the scene is a conscious behaviour such that hypotheses and expectations of the moving patterns of objects being observed are made constantly [5]. It is for such reasons that we address in this work the problem of how moving objects in a scene can be observed with expectations in order to provide cues for selective visual attention. Similar work has been addressed by [2, 9, 10].

Meaningful objects always move with purposes. In a known environment, such inherent purposes are associated with patterns of moving sequences which are constrained by the spatio-temporal characteristics of the environment. Moving purposes of an object is often distinctively captured by the spatio-temporal regularities in its movement patterns. A common phenomena would be the observation of service vehicles entering an airport docking stand (figure 1). The hidden regularities can be regarded as a set of conditional dependencies in space and time and such spatio-temporal dependencies are mostly qualitative and probabilistic. Inspired by Rimey and Brown's recent study in active vision [8], we extend the use of Augmented Hidden Markov Model to model the probabilistic spatio-temporal regularities in object's movement for providing visual expectations and selective visual attention.

2 "Blind" Expectations

Hidden Markov Model (HMM) has been widely used in speech recognition for modelling and classifying sound patterns. A HMM is essentially defined by its initial state probability distribution π, the symbol probability distribution B and the state transition probability distribution A. A HMM represents the probabilistic characteristics of a sequential pattern in two levels: (1) a state sequence represents a sequential combination of "hidden" hypotheses; (2) an observation symbol sequence models the most likely combination of local evidence for the transitions between the states. A detailed overview of HMM can be found in [7]. In figure 1, when a vehicle enters the scene, the spatial locations at which significant changes in orientation of vehicle's movement occur are the states. The visual evidence, orientation and displacement of the vehicle's movement, are the symbols. Assuming that there is only a very weak correlation between the orientation and the displacement of vehicle's movement, we can use a pair of independent HMMs to model the orientations and displacements simultaneously. In other words, a HMM can be applied to capture the probabilistic moving regularities of a type of vehicle. The probability distributions of a HMM are learned from examples. Provided with training sequences, multiple HMMs can be established for giving probabilistic spatio-temporal expectations of different vehicle appearances (see figure 2).

The essence in HMM learning can be characterised as a process of establishing impacts of updated visual evidence from the training sequence on the model's partial conditional beliefs [1]. More precisely, to compute: *(1) the conditional probabilities of the partial observation sequence* $O_1, O_2, ..., O_t$ *and state to be in*

*The work is funded by the ESPRIT EP2152 (VIEWS) project.

[1] An excellent systematic analysis of graph-based belief networks can be found in [6].

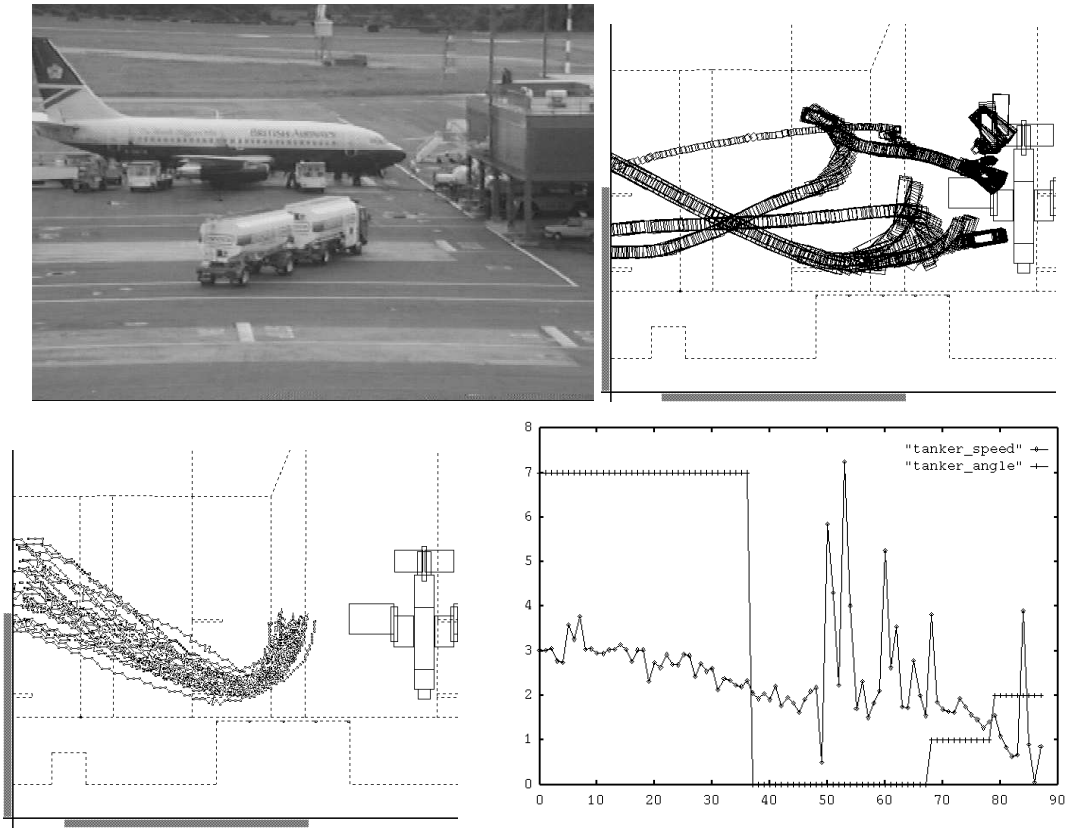

Figure 1: Moving patterns of service vehicles at an airport docking stand. (1) a fuel tanker and a trailer in service; (2) moving patterns of typical service vehicles in the ground plane; (3) 30 training sequences for the fuel tanker and (4) overlapped discrete orientations and the associated frame-wise displacements of the tanker.

q_i *at time* t*, given the model*,

$$\alpha_t(i) = [\sum_{j=1}^{N} \alpha_{t-1}(j)a_{ji}]b_i(O_t) \tag{1}$$

where $\alpha_1(i) = \pi_i b_i(O_1)$, $1 \le i \le N$, $2 \le t \le T-1$; *(2) the conditional probabilities of the partial observation sequence* $O_{t+1}, O_{t+2}, ..., O_T$*, given the state had been in* q_i *at* t *and the model*,

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}b_j(O_{t+1})\beta_{t+1}(j) \tag{2}$$

where T is the length of the sequence, $\beta_T(i) = 1$ and $1 \le i \le N$, $t = T-1, T-2, ..., 1$. Based on equations (1) and (2), *Baum-Welch* learning algorithm [7] adjusts the probability distributions for a single learning observation sequence. However, a reliable estimate of HMM λ can only be obtained through multiple learning sequences. Let us denote a set of K learning examples as $\mathbf{O} = [\mathbf{O}^{(1)}, \mathbf{O}^{(2)}, ..., \mathbf{O}^{(K)}]$, where $\mathbf{O}^{(k)} = [O_1^{(k)}, O_2^{(k)}, ..., O_{T_k}^{(k)}]$ is the kth sequence. If each sequence is independent of all others, then we should have:

$$\bar{a}_{ij} = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i)a_{ij}b_j(O_{t+1}^{(k)})\beta_{t+1}^k(j)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i)\beta_t^k(i)},$$

$$\bar{b}_j(l) = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} d_t\alpha_t^k(i)\beta_t^k(i)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i)\beta_t^k(i)} \tag{3}$$

where

$$d_t = \begin{cases} 1 & \text{if } O_t = V_k \\ 0 & \text{otherwise,} \end{cases}$$

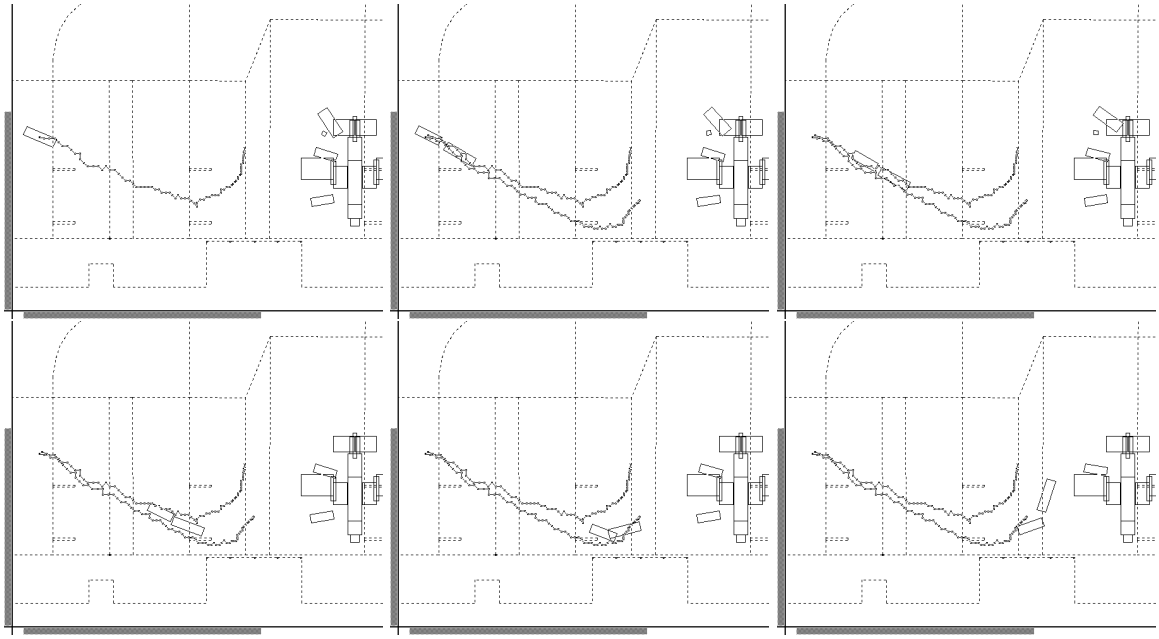

Figure 2: The appearance of a fuel tanker initialises a visual blind expectation of future movement of that vehicle. After 9 frames, a similar blind expectation was caused by the appearance of a fuel trailer. The next 4 ground plane diagrams show the positions of the vehicles at every 20 frames interval overlapped on the initial expectations.

$$P_k = \quad P[\mathbf{O}^{(k)}|\lambda] = \quad \sum_{i=1}^{N} \alpha_{T_k}(i) \tag{4}$$

3 Visual Expectations

Visual observation can be regarded as an ongoing process of adjusting our underlying expectations of object behaviour with instantaneous updated visual evidence and simultaneously applying such modified expectations to guide the visual perception in order to guarantee the effectiveness and correctness of future visual evidence (see figure 3).

Visual augmentation on HMM was introduced for foveation control in active visual sensing [8]. The concept extends naturally to visual observation. Now, consider that: (1) the partial conditional beliefs are functions of time, i.e. we have $\lambda(t)$ with a particular value at time t denoted as λ^t; and (2) the updated visual evidence of the moving object is regarded as the immediate prediction of the $\lambda(t)$'s symbol output. That is, if i is the index of the state q_i determined by λ^t at current time t, k is the index of the observation symbol v_k at time t, and assuming that λ^{t+1} will produce v_k at time $t+1$, then a belief modification weight is [8]:

$$\omega_j^t = \frac{a_{ij}^t b_j^t(k)}{\sum_{l=1}^{N} a_{il}^t b_l^t(k)}, \qquad 1 \le j \le N \tag{5}$$

ω_j^t represents a conditional belief that $\lambda(t)$ will be in state q_j at time $t+1$, given λ^t is in state q_i and has received visual evidence for output symbol v_k at time $t+1$. Assuming that symbol v_k will be generated by λ^{t+1}, this conditional belief weighting factor gives approximately the state transition probability $\tilde{a}_{ij}^{t+1}$ at time $t+1$, i.e. $\tilde{a}_{ij}^{t+1} = \omega_j^t$ where $1 \le j \le N$ [2]. Once again, applying the same assumption that current visual evidence v_k will be the immediate future output symbol of $\lambda(t)$, and considering that this conditional belief is also dependent on the probability of $\lambda(t)$ being in a particular state, we have:

$$\tilde{b}_j^{t+1}(l) = \frac{w_j^t d_j^t(l)}{\sum_{m=1}^{M} w_j^t d_j^t(m)}, \quad 1 \le l \le M \tag{6}$$

[2] A detailed analysis can be found in [8].

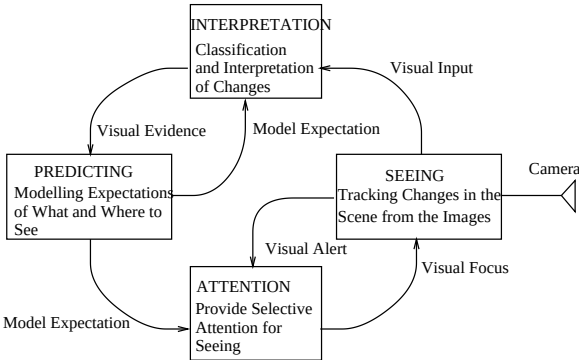

Figure 3: A "see-predict-see" feedback loop in visual observation.

where

$$d_j^t(l) = \begin{cases} 1 & \text{if } l = k \\ 0 & \text{otherwise} \end{cases}$$

Further, considering that changes to the a_{ij}^t and $b_j^t(l)$ should be taken gradually, therefore controlled by a modification gain, and such changes can only be maintained if the same visual evidence is maintained, i.e. controlled by a decay gain, then:

$$a_{ij}^{t+1} = r_{dg}[r_{mg}\omega_j^t + (1 - r_{mg})a_{ij}^t] + (1 - r_{dg})a_{ij}^t \quad (7)$$

$$b_j^{t+1}(l) = s_{dg}\{\frac{s_{mg}w_j^t d_j^t(l) + (1 - s_{mg})b_j^t(l)}{\sum_{m=1}^{M}[s_{mg}w_j^t d_j^t(m) + (1 - s_{mg})b_j^t(m)]}\} + (1 - s_{dg})b_j^t(l) \quad (8)$$

where r_{mg} and s_{mg} $(0 \leq r_{mg}, s_{mg} \leq 1)$ are state and symbol modification gains, r_{dg} and s_{dg} $(0 \leq r_{dg}, s_{dg} \leq 1)$ are state and symbol decay gains respectively. As there is no visual evidence for the initial time step, the initial state distribution π cannot be adjusted.

4 Experiments and Discussions

Figure 2 shows blind expectations after the appearances of a fuel tanker and a fuel trailer have been detected. The length of the expectation is determined by the probability $P[O|\lambda_i]$ given by equation (4). Figure 4 shows the visually augmented expectations. It is evident that the expectations for the immediate future 20 frames are more accurate than for the distant future as the instantaneous visual evidence only influences expectations locally in time. It is also evident that under the current circumstance, no direct effect exists on the visual tracking from which visual evidence is provided. An assumption was made that visual evidence at each time frame was collected under the guidance of the expectation. Our immediate future work will be on establishing this visual feedback link illustrated in figure 3.

We described the need to have visual expectations for selective attention in visual observation and, more importantly, that selective attention should be context dependent. We illustrated how Augmented Hidden Markov Models can be used to capture the intrinsic spatio-temporal regularities of moving vehicles in a known scene and consequently, to predict vehicle's movement with instantaneous visual augmentation. In this airport scenario, scripts are used to describe the expected vehicle service steps in loading and delivering. By linking the discrete "hidden" states of AHMM to the steps in the scripts, we should be able to deliver meaningful conceptual descriptions of the vehicle behaviour.

References

[1] D. Ballard. Reference Frames for Animate Vision. In *IJCAI*, Detroit, USA, Aug. 1989.

[2] P.J. Burt. Smart Sensing in Machine Vision. In *Machine Vision: Algorithms, Architectures and Systems*. Academic Press, San Diego, Ca., 1988.

[3] J.J. Gibson. *The Perception of the Visual World*. Greenwood Press, Westport, CT, USA, 1950.

[4] I.E. Gordon. *Theories of Visual Perception*. John Wiley & Sons, New York, 1989.

[5] H. von Helmholtz. *Popular Scientific Lectures*, chapter The Recent Progress of the Theory of Vision. Dover Publications, New York, USA, 1962.

[6] J. Pearl. *Probabilistic Reasoning in Intelligent Systems, Networks of Plausible Inference*. Morgan Kaufman Publ. Inc., Los Altos/CA, 1988.

[7] L.R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[8] R.D. Rimey and C.M. Brown. Selective Attention as Sequential Behavior: Modeling Eye Movements with an Augmented Hidden Markov Model. In *DARPA Image Understanding Workshop*, pages 840–850, Pittsburgh, USA, Sept. 1990.

[9] T.M. Sobh and R. Bajcsy. Visual Observation as a Discrete Event Dynamic System. In *IJCAI Workshop on Dynamic Scene Understanding*, Sydney, Australia, Aug. 1991.

[10] J.K. Tsotsos. On the Relative Complexity of Active vs. Passive Visual Search. *International Journal of Computer Vision*, 7(2):127–141, 1992.

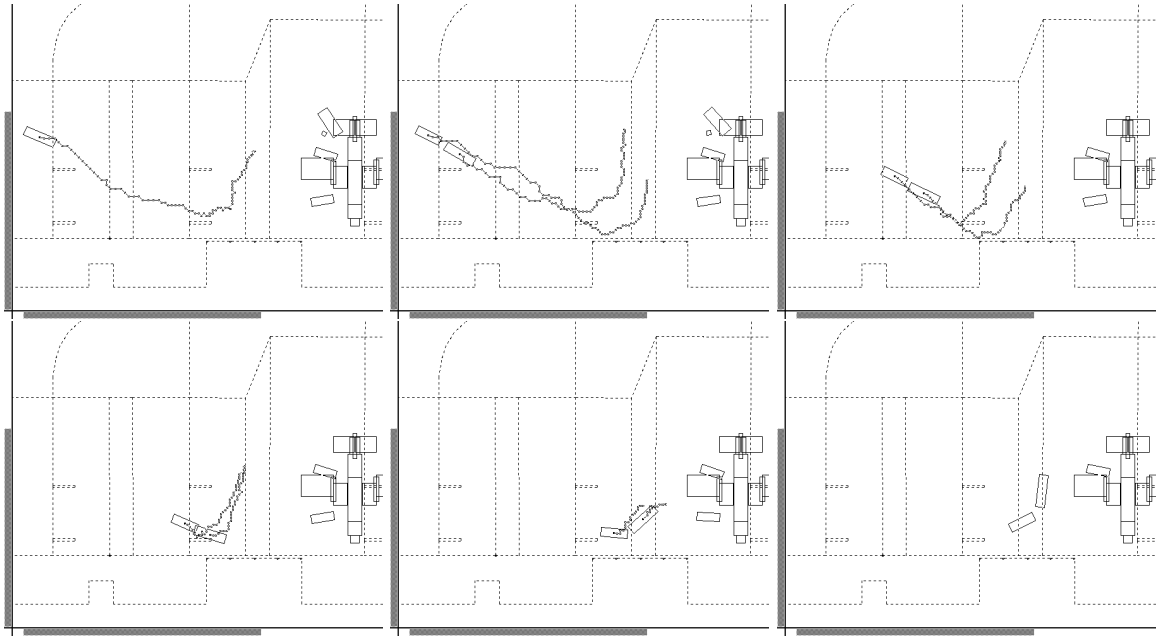

Figure 4: The same scenario as in figure 2 but with visual augmentation at each frame. The gains from the visual augmentation are $s_{mg} = 0.1$, $s_{dg} = 0.1$, $r_{mg} = 0.4$ and $r_{dg} = 0.4$. The state transition gains should always be much smaller compared with the observation symbol gains because the instantaneous visual evidence at each frame should not have strong influence on the changes in its movement pattern. Also, because the current see--predict-see loop still lacks the "Visual Focus" and "Model Expectation" links shown in figure 3. Therefore, the confidence in the visual evidence under the current circumstance is low and all the gains for the augmentation have been set low.