Visual observation as reactive learning

Shaogang Gong

Computer Science Department, Queen Mary and Westfield College
Mile End Road, London E1 4NS, England

ABSTRACT

Meaningful objects in a scene move with purpose. The ability to induce visual expectations from such purpose is important in visual observation. By regarding the spatio-temporal regularities in the moving patterns of an object in the scene as a network of temporally dependent belief hypothesis, visual expectations can be represented by the most likely combinations of the hypotheses based on updating the network in response to instantaneous visual evidence. A particular type of probabilistic single path Directed Acyclic Graph (DAG) belief network, the Hidden Markov Model (HMM), can be used to represent the "hidden" regularities behind the apparently random moves of an object in a scene and reproduce such regularities as "blind", therefore, insensitive expectations. By adaptively adjusting such a probabilistic belief network with observed visual evidence instantaneously, a Visual Augmented Hidden Markov Model (VAHMM) can be used to model and produce dynamic expectations of a moving object in the scene. In particular, using tracked moving service vehicles at an airport docking stand as visual cues, we present how a VAHMM can be constructed first to represent the probabilistic spatial dependent relationships in the typical moving patterns of a type of vehicle, and then to adjust the weighting parameters of such dependencies dynamically with instantaneous new visual evidence. We describe the use of such model to generate in time the probabilistic expectations of an observed object and discuss some possible initial applications of such a framework for providing selective attention in visual observation.

1. INTRODUCTION

Visual observation of moving objects for understanding dynamic scene has been studied extensively in computer vision [6, 9, 7, 10, 11]. However, most approaches have so far ignored the use of any knowledge about the scene and about the objects being observed. Consequently, the complexity involved in tracking models in two-dimensions or three-dimensions in such an indiscriminate manner reveals that even massive parallelism cannot overcome sufficiently the computational burden required for real time dynamic scene understanding [19]. In fact, visual processing is highly selective, purposive and active [8, 17, 2, 1], whether it is for providing cues in a decision making process for accomplishing given tasks or for observing, understanding, and interpreting changing world. Task knowledge and the nature of the scene often define the visual attention and allow us to ignore the irrelevant [8].

In active vision, visual perception is guided constantly by the intentions of a decision making process and used to provide information for accomplishing such intentions. Mechanisms for attentional focus for visual processing have been studied and various frameworks have been proposed [1, 5, 16, 20]. However, visual observation of dynamic scenes in computer vision has still been treated merely as a passive process and such a purposive concept in active visual sensing has not been widely applied. In fact, whether visual observation is for the sake of understanding and interpreting the scene or merely for "watching", it is a conscious behaviour such that hypotheses and expectations of the spatio-temporal regularities in the moving patterns of objects being observed are made adaptively according to the changes in the scene. It is for such reasons that we address in this work the problem of how the inherent purposes of moving objects being observed in a scene can be modeled dynamically in order to provide cues for selective attention in machine visual observation. Similar work for "smart" visual observation has been addressed by [3, 4, 18].

In the following, we first argue that "hidden" intentions in an object's movement in a known scene can be defined by the spatio-temporal regularities in its moving patterns and such regularities can be modelled appropriately by probabilistic belief networks, in particular, the Hidden Markov Model. Then we address the issue of collecting visual evidence

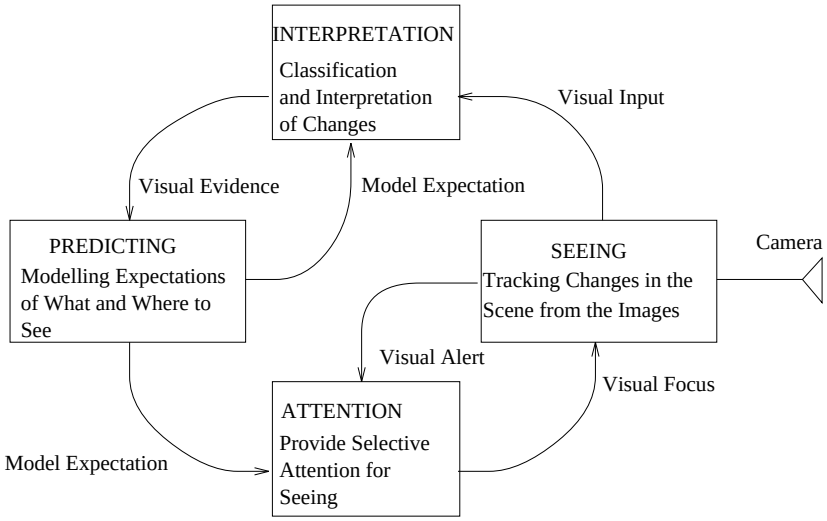

Figure 1: A "see-predict-see" feedback loop in visual observation.

and estimate its impact on a Hidden Markov Model in order to provide a "see-predict-see" close feedback loop in visual observation (see figure 1). We also present how such visual augmentation can be performed by applying a Visual Augmented Hidden Markov Model. We discuss some aspects of how a see-predict-see loop can be established and finally, address some of our immediate and long term future work.

2. MODELLING OBJECT'S MOVING PATTERNS

Meaningful objects always move with purposes. In a known environment, such inherent purposes appear as different patterns of moving sequences that are associated with certain "hidden" regularities which are constrained by the spatio-temporal characteristics of the environment. It is feasible that the moving purposes of an object is distinctively captured and can be defined by the different spatio-temporal regularities in its movement patterns. For example, when observing a person who is walking into our laboratory, we are able to "guess" his next possible moves, i.e. make hypotheses, with some degree of uncertainty, and watch him with anticipation. By continuing our observation, our "guesses" on his movement will become more certain and our understanding of his intention will become clearer. Similar phenomena would be the observation of service vehicles entering an airport docking stand (figure 2). 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. Attempts to model them based on deterministic geometric functionals with optimisation procedures is perhaps overcommitted and therefore may be inappropriate. On the other hand, a directed graphic probabilistic belief network captures the essence of such dependent relationships [14] and can be exploited for modelling phenomena of such nature. Inspired by Rimey and Brown's recent study in active vision [16], we extend the use of Hidden Markov Model, one kind of probabilistic Directed Acyclic Graph (DAG) belief network, to the representation of probabilistic spatio-temporal regularities of moving objects.

2.1. Hidden Markov Model

Hidden Markov Model (HMM) has been widely used in speech recognition for modelling and classifying sound patterns. Despite its well understood ability to represent spatio-temporal regularities of conditional dependencies in sequential patterns, its use in computer vision had been hardly exploited until recent work in active vision by Rimey and Brown [16], in which an augmented Hidden Markov Model was proposed for modelling the foveation path of an active head. In this work, we extend such use of HMM to the modelling of spatio-temporal dependencies in an object's movement. A detailed overview of HMM can be found in [15]. In the following, we briefly summarise some essential characteristics of HMM.

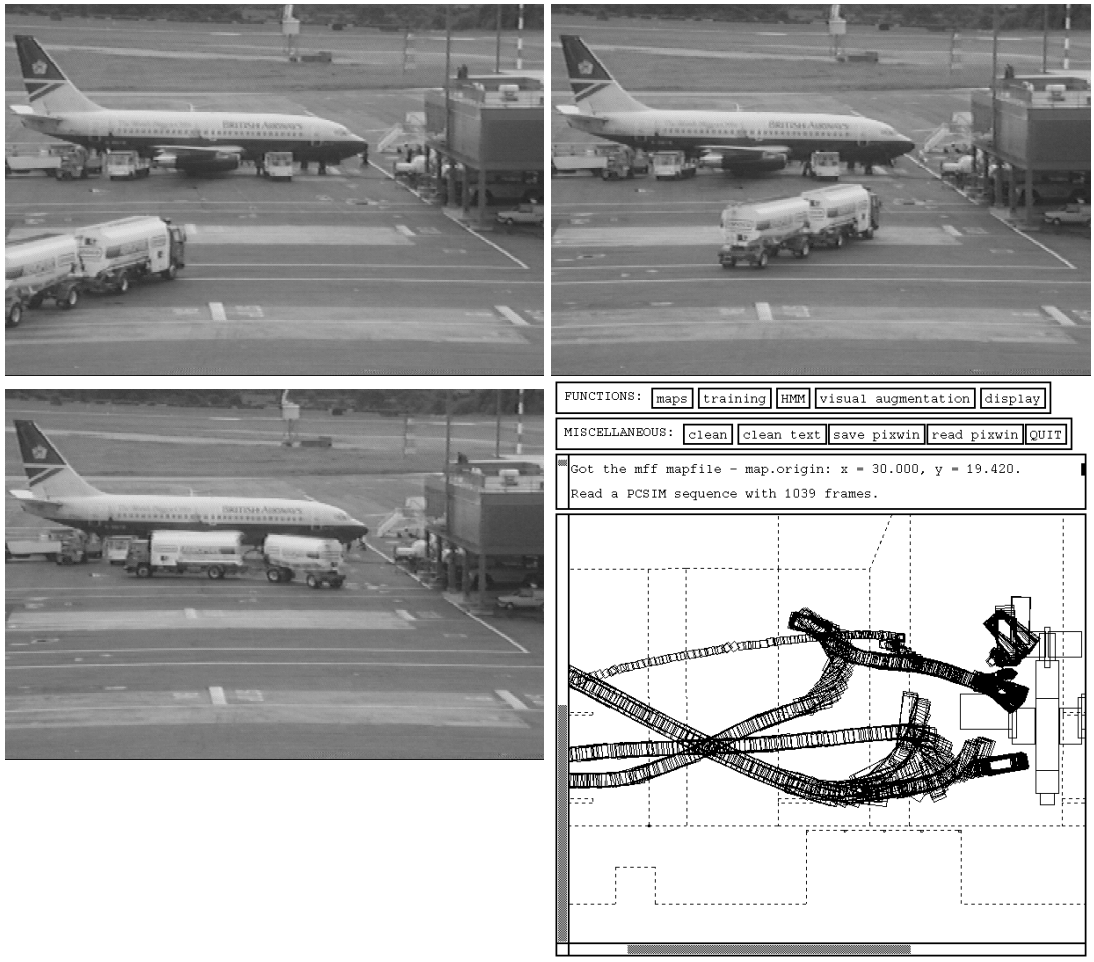

Figure 2: An example of moving patterns of service vehicles at an airport docking stand. The first three images show different stages of a fuel tanker and a fuel trailer in their service path, at where indicate the most likely entering, maneuvering and stopping positions for such service vehicles in this airport stand scenario. The last diagram illustrates moving patterns of 6 types of most common service vehicles in the ground plane of this airport stand. The x and y axes of the scene coordinate are respectively along the white line in front of the building and from the right to the left in the images. The x and y axes in the ground plane are respectively from the left to the right and from the bottom to the top in the diagram shown. The scales of both coordinates are in metres.

HMM is regarded as a probabilistic model of causal dependencies between different status in sequential patterns and a special case of Bayesian belief network [14]. In discrete form, it can also be regarded as a stochastic finite state network [13]. The parameters of a HMM are learned by past examples. The model then can be used to classify a given sequential sequence or to generate sequences that inherit the causal dependencies between successive steps in the sequence and therefore can be regarded as expectations of most likely combinations of hypothesis. A HMM represents the probabilistic characteristics of a sequential pattern at two levels: 1) first, a state sequence, which represents a sequential combination of "hidden" hypotheses under the current probabilistic distribution of the state dependencies; 2) second, an observation symbol sequence, which models the most likely combination of local evidence, i.e. apparent visual causes, for the transitions between the states. In the airport scenario, when a vehicle enters the scene, the spatial locations at where significant changes in orientation of vehicle's movement occur are defined as the states, and the visual causes, orientation and displacement of the vehicle's movement, are taken as 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 simultaneous. Figure 3 shows a typical *Bakis*, also known as the

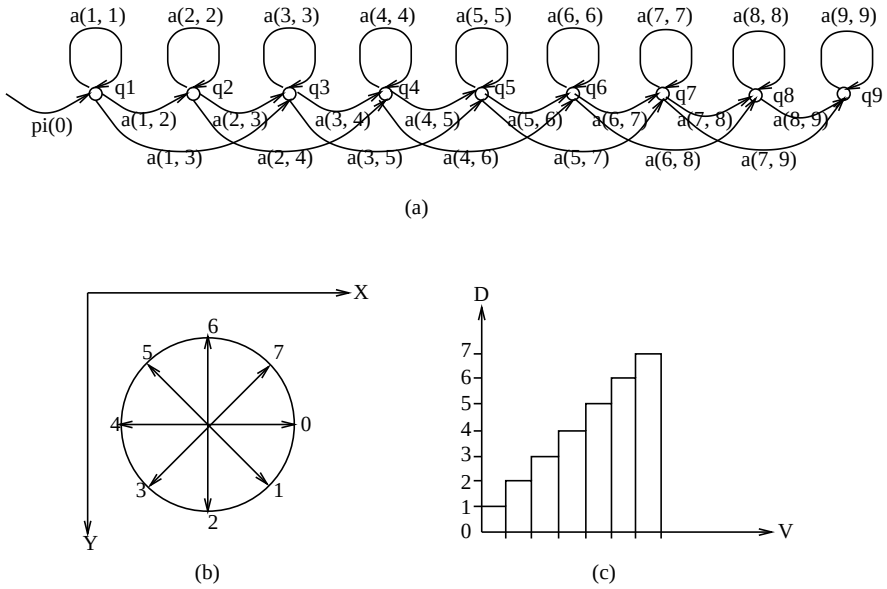

Figure 3: (a) A typical *Bakis* HMM with probability distribution associated with the state transitions. (b) 8 discrete orientations in the ground plane are taken as one set of observation symbols associated with object's moving patterns. (c) 8 discrete displacements which associate with the speed variations in object's movement in the ground plane are taken as another set of observation symbols.

left-right HMM, that has symbols as 8 discrete orientations and displacements between two successive frames. Their values are learned from past examples and vary between different type of vehicles.

A HMM is fully given by the following factors:

- the number of defined states N and the defined states $Q = \{q_1, q_2, ..., q_N\}$;
- the number of defined symbols M and the defined symbols $V = \{v_1, v_2, ..., v_M\}$;
- the state transition probability distribution $A = \{a_{ij}\}$, where $a_{ij} = P[q_j \text{ at } t+1 | q_i \text{ at } t]$, $1 \leq i, j \leq N$.
- the symbol probability distribution $B = \{b_j(k)\}$, where $b_j(k) = P[v_k \text{ at } t | q_j \text{ at } t]$, $1 \leq j \leq N$, $1 \leq k \leq M$.
- initial state distribution $\pi = \{\pi_i\}$, where $\pi_i = P[q_i \text{ at } t = 1]$, $1 \leq i \leq N$.

With defined Q and V, given M, N, A, B and π, a HMM is noted as $\lambda(A, B, \pi)$. At any discrete time, a λ will always be in one state with a particular symbol according to the probability distributions of A, B and π. Therefore, these parameters of λ are weighting factors that describe the strength of the dependencies between the states and between the states and symbols. They represent local conditional beliefs that their combined effect gives very likely combinations of hypothesis in sequences.

2.2. Learning object's moving patterns with HMM

Service vehicles in airport docking stands move with purpose and their moving patterns always associate with some spatio-temporal regularities. Such regularities are most likely probabilistic as illustrated by the patterns in figure 2. A HMM can be applied to capture the regularities of a type of vehicle by assigning its state hypotheses as the most likely "spots" in the ground plane at where significant changes in orientation of the movement occur, and by corresponding its symbol evidences to the instantaneous orientation and displacement of vehicle's movement.

There are four essential uses for HMM:

1. Classification: for a given observation symbol sequence $O = \{O_1, O_2, ..., O_T\}$, where T is the length of the sequence, by computing $P[O|\lambda_i]$ for a set of known λ_i, the sequence can be classified according to λ_i where Max $P[O|\lambda_i]$ occurs.

2. Explanation: given $O = \{O_1, O_2, ..., O_T\}$ and a HMM λ, by applying the *Viterbi* algorithm [15], a single most likely state sequence $Q = \{q_1, q_2, ..., q_T\}$ can be found.

3. Learning: given an example observation sequence $O = \{O_1, O_2, ..., O_T\}$ and a model λ, the model parameters of λ can be adjusted such that $P[O|\lambda]$ is maximised.

4. Generation: Given a λ, it can be used as a generator to produce the observation symbol sequences and their associated state sequences in which the probabilistic characteristics of the model are inherently reflected.

The work reported here is mainly related to the issues of learning the models from a set of training sequences and of generating future observation sequences (see figure 4).

The essence in learning a λ can be characterised as a process of establishing impacts of each updated visual evidence from the learning sequence on the model's partial conditional beliefs, i.e. the probability distributions of the model, by forward and backward propagations along the model's graph network [1]. More precisely, if $O = \{O_1, O_2, ..., O_T\}$, then compute:

- the conditional probabilities of the partial observation sequence $O_1, O_2, ..., O_t$ and state to be in q_i at time t, given the model λ,

$$\alpha_t(i) = P[O_1, O_2, ..., O_t,\ q_i \text{ at } t|\lambda] = [\sum_{j=1}^{N} \alpha_{t-1}(j)a_{ji}]b_i(O_t), \qquad 1 \le i \le N,\ 2 \le t \le T-1$$

  where $\alpha_1(i) = \pi_i b_i(O_1)$;

- 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 λ,

$$\beta_t(i) = P[O_{t+1}, O_{t+2}, ..., O_T|q_i \text{ at } t,\ \lambda] = \sum_{j=1}^{N} a_{ij}b_j(O_{t+1})\beta_{t+1}(j), \qquad 1 \le i \le N,\ t = T-1, T-2, ..., 1$$

  where T is the length of the sequence and $\beta_T(i) = 1$.

After some manipulations on the *Baum-Welch* learning algorithm described in [15], we can then adjust, for a single learning observation sequence, the parameters of λ by:

$$\tilde{\pi}_i = P[q_i \text{ at } 1|O,\ \lambda] = \frac{\sum_{j=1}^{N} P[q_i \text{ at } 1,\ q_j \text{ at } 2,\ O|\lambda]}{P[O|\lambda]} = \frac{\alpha_1(i)\beta_1(i)}{\sum_{i=1}^{N} \alpha_1(i)\beta_1(i)} \qquad 1 \le i,\ j \le N \tag{1}$$

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T-1} P[q_i \text{ at } t,\ q_j \text{ at } t+1|O,\ \lambda]}{\sum_{t=1}^{T-1} P[q_i \text{ at } t|O,\ \lambda]} = \frac{\sum_{t=1}^{T-1} \alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} \tag{2}$$

$$\tilde{b}_j(k) = \frac{\sum_{t=1}^{T} d_t P[q_j \text{ at } t|O,\ \lambda]}{\sum_{t=1}^{T} P[q_j \text{ at } t|O,\ \lambda]} = \frac{\sum_{t=1}^{T} d_t \alpha_t(j)\beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)} \qquad 1 \le k \le M \tag{3}$$

[1] An excellent systematic analysis of graph-based belief networks can be found in [14].

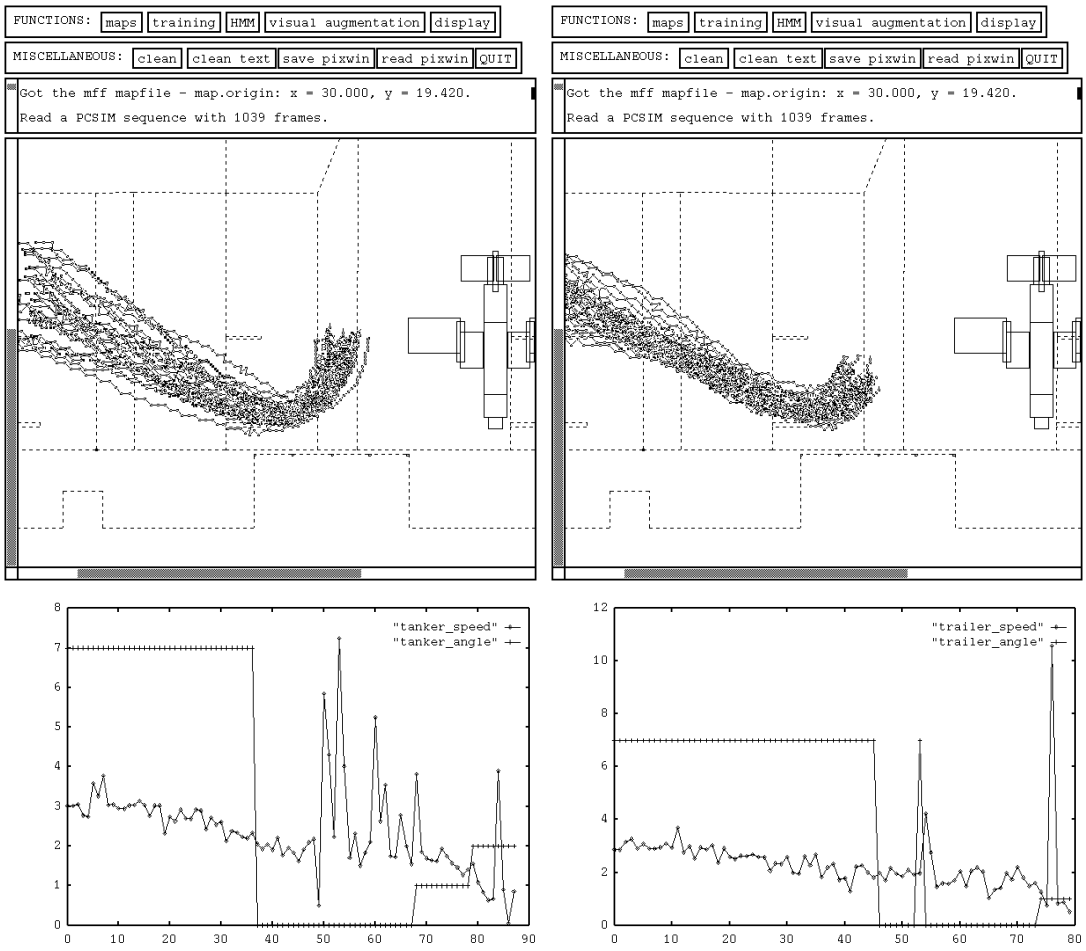

Figure 4: Training set of 30 sequences generated based on uniform distribution of random variations on an example moving sequence of the fuel tanker and trailer on the ground plane. The plotting diagrams show the overlapped discrete orientations and the associated frame-wise displacements in the moving sequence. The x axis is in time frames and y axis is between 0 - 7 for the orientations and in metres for the displacement.

where

$$P[q_i \text{ at } t|O,\lambda] = \frac{\sum_{j=1}^{N} P[q_i \text{ at } t,\, q_j \text{ at } t+1,\, O|\lambda]}{P[O|\lambda]} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}, \qquad d_t = \begin{cases} 1 & \text{if } O_t = V_k \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Denominators in equations (1), (2), (3) and (4) are all normalisation constants. A reliable estimate of a model 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, it is straightforward for us to have the following equations for learning the model from multiple sequences:

$$\bar{a}_{ij} = \frac{\sum_{k=1}^{K}\frac{1}{P_k}\sum_{t=1}^{T_k-1}\alpha_t^k(i)a_{ij}b_j(O_{t+1}^{(k)})\beta_{t+1}^k(j)}{\sum_{k=1}^{K}\frac{1}{P_k}\sum_{t=1}^{T_k-1}\alpha_t^k(i)\beta_t^k(i)}, \qquad \bar{b}_j(l) = \frac{\sum_{k=1}^{K}\frac{1}{P_k}\sum_{t=1}^{T_k-1} d_t\alpha_t^k(i)\beta_t^k(i)}{\sum_{k=1}^{K}\frac{1}{P_k}\sum_{t=1}^{T_k-1}\alpha_t^k(i)\beta_t^k(i)} \tag{5}$$

where

$$P_k = \quad P[\mathbf{O}^{(k)}|\lambda] = \quad \sum_{i=1}^{N} \alpha_{T_k}(i) \tag{6}$$

and d_t is given by equation (4). Now, a *Bakis* HMM with a graph of characteristics similar to figure 3 (a) is in fact not necessary to learn π_i since its $\pi_1 = 1$ and $\pi_i = 0$ where $i \neq 1$. It is likely that most of the sequential behaviours with which we are concerned can be dealt with by using the *Bakis* model. Applying the above equations to a set of training sequences, a vehicle's HMM can be trained iteratively until a minimum error threshold is met. With different training sets for different types of vehicles, a set of HMM λ can be established for producing spatio-temporal expectations of vehicle movement whenever there is an appearance of any of the vehicles that has a trained model (see figure 5).

Usually, observation sequence generation is based on maximum likelihood state transition and symbol output of the λ at each time step, in which case an unique expectation sequence will always be produced by the same model. However, Rimey and Brown [16] pointed out that by generating the maximum likelihood sequence, one is bound to the assumption that knowledge of the moving object and of the scene has not been changed since the example sequences were taken and will not change in future. It is quite obvious that such an assumption is inadequate since, although the model has learned the dynamic characteristics of the object based on past observations, the behaviour of a current moving object could still change according to the current conditions in the scene. Thus, the model should be flexible and able to adjust. Such flexibility in a model can be introduced by giving a degree of randomness in the process of sequence generation, i.e. instead of generating the maximum likelihood sequence, state transitions and symbol outputs at each time step could be based on a unified random selection in the appropriate probability distribution of the model [2]. Based on our own experiments, although such random generation is desirable for symbol outputs, it is too sensitive and unstable in state transitions as any randomly triggered early transition will cause the sequence to wander away without reflecting the intrinsic nature of the object's movement. Consequently, a mixed random symbol output with a durational controlled [3] state transition [15] seems to be more appropriate.

3. VISUAL AUGMENTATION FOR DYNAMIC LEARNING

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. In other words, it is a process of reactive learning (see figure 1).

It is desirable that this two-way feedback effect is represented in the HMM. Visual augmentation on HMM was introduced for foveation control in active visual sensing [16]. The concept extends naturally to visual observation, although it may be used in a different context. By forming a weighted sum of various visual inputs, e.g. detected moving patterns in the image plane or projected moving targets in the ground plane by various methods, an updated visual evidence of the moving object in the scene can be represented. 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 given by [16] as:

$$\omega_j^t = P[q_j \text{ at } t+1 | q_i \text{ at } t,\, v_k \text{ at } t+1] = \frac{a_{ij}^t b_j^t(k)}{\sum_{l=1}^{N} a_{il}^t b_l^t(k)} \qquad 1 \leq j \leq N \tag{7}$$

[2] Personal communications with Ray Rimey.

[3] It is better known as "semi-hidden" Markov models in speech recognition.

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 $1 \leq j \leq N$ [4]. 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:

$$\tilde{b}_j^{t+1}(l) = \frac{w_j^t d_j^t(l)}{\sum_{m=1}^{M} w_j^t d_j^t(m)} \qquad 1 \leq j \leq N, \quad 1 \leq l \leq M, \quad d_j^t(l) = \begin{cases} 1 & \text{if } l = k \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The denominator in equation (8) is a normalisation factor to ensure $\sum_{l=1}^{M} \tilde{b}_j^{t+1}(l) = 1.0$. Now, by 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 also considering such changes can only be maintained if the same visual evidence is maintained, i.e. controlled by a decay gain, the final modifications to a_{ij}^t and $b_j^t(l)$ should be:

$$a_{ij}^{t+1} = r_{dg}[r_{mg}\omega_j^t + (1 - r_{mg})a_{ij}^t] + (1 - r_{dg})a_{ij}^t \tag{9}$$

$$b_j^{t+1}(l) = s_{dg}\{\frac{s_{mg} w_j^t d_j^t(l) + (1 - s_{mg})b_j^t(l)}{\sum_{m=1}^{M}[s_{mg} w_j^t d_j^t(m) + (1 - s_{mg})b_j^t(m)]}\} + (1 - s_{dg})b_j^t(l) \tag{10}$$

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, i.e. the first visual input is taken as the visual evidence for the next time step, the initial state distribution π cannot be adjusted and this reflects the characteristic of a Bakis model as we described earlier. Equations (8), (9) and (10) are derived by some simple manipulations on the results given in [16].

Taking the above equations as reactive modifications to the partial conditional beliefs that coordinate to govern a global belief and expectation for the object's movement in the scene, VAHMM can be used to produce dynamically an updated prediction of how an ongoing moving object will appear in the immediate future. Such procedure takes three basic steps in a loop and echos the see-predict-see feedback loop (figure 1) discussed earlier:

1. $\lambda(t)$ generates a sequence of $O_t, O_{t+1}, ..., O_T$ and a sequence of $S_t, S_{t+1}, ..., S_T$ at time t;
2. these future symbol and state sequences are used to guide the observation and consequently, a new visual evidence is collected at time t;
3. this visual evidence at time t is used to augment the λ^t and to produce λ^{t+1}. Then increase the time step and go back to step one.

4. EXPERIMENTS

The VAHMM approach for visual observation as reactive learning has been tested at the airport service stand scenario (see figure 2). With the frame rate of 2.5 hertz, a three-dimensional model-based tracking process [12] is applied to detect and track any appearance and movement that are identified with pre-classified models. Figure 5 shows the initial 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 (6). The threshold of $P[O|\lambda_i]$ is set small currently in order

[4] A detailed analysis can be found in [16].

to provide a longer valid expectation (89 future frames for the tanker and 79 future frames for the trailer). However, small threshold causes greater uncertainty in the distant future expectations. This is evident in figure 5.

For the same scenario, figure 6 shows the visually augmented expectations of the fuel tanker and fuel trailer. It is evident that the expectations for the immediate 20 future frames are more accurate than for the distant future as the instantaneous visual evidence influences the model dynamically. This is because that the visual influence is only "local" to the model belief dependency and it stays short as the decay gains have been set low (see reasons stated in figure 6). It is also evident that under the current circumstance, no direct effect exists on the visual tracking from which visual evidence is provided, i.e. 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 concentrated on establishing this visual feedback link illustrated in figure 1.

5. SUMMARY AND FUTURE WORK

We described the need to have selective attention in visual observation and, more importantly, that selective attention should be context dependent. In particular, we propose that it can be modelled by a probabilistic belief network of hypothesis which reflects the "hidden" purposes in the movement of the object being observed. The causal dependencies in the network could be extracted from the apparent moving patterns of the object. The work described here presents some initial attempts of our long term study for exploiting an appropriate mechanism that provides selective attention in machine vision observation at real time. We illustrated how one specific extended graph belief network, the VAHMM, developed for active visual sensing, can be used to model the intrinsic spatio-temporal regularities of dynamic objects in a known scene and consequently, to predict object's movement with instantaneous visual augmentation.

Our ongoing and immediate future work will be concentrated on testing such VAHMM on multiple moving objects and establishing selective attention in the visual tracking process. Also, in terms of using VAHMM, in addition to learning and generation of expectations, we will also exploit explanation of observed sequences and link the discrete states to vehicle scripts in order to give conceptual descriptions of vehicle behaviour.

One of the immediate extension of the use of VAHMM will be to set dynamically the tracking threshold and reduce searching domain in model matching. The current model tracking in VIEWS (ESPRIT EP2152 project "Visual Inspection and Evaluation of Wide-area Scenes") uses a Kalman filter to continuously update feature matches in image sequence [12]. However, with selective attention, we may be able to trade off resources spent on the model matching where the trajectory is predictable using feedback from the discrete states of the VAHMM.

Another use of the model is to provide missing gaps in visual evidence, such as in a case of occlusion. Also, by establish multi-VAHMMs for multi-type objects, we constantly concentrate visual observation on the meaningful targets and ignore the irrelevant. Setting such processing priority is mediated by high level knowledge of the domain. In the current VIEWS airport stand scenario, we use scripts to describe the expected vehicle path in terms of the service steps in loading and delivering. By linking the discrete "hidden" states of VAHMM to the steps of the scripts, we are able to deliver status reports and meaningful conceptual descriptions of the vehicle behaviour for end users.

VAHMM as presented has some limitations. First, it is rather unclear how visual evidence that has been confirmed for long duration can permanently alter a particular partial conditional belief learned from training examples. Second and more fundamentally, the model is incapable of adding new transition links between states, nor is it possible to add new states. From our early study in the more general probabilistic belief networks, it is evident that Bayesian belief networks [14] may provide an alternative to overcome these limitations.

6. ACKNOWLEDGEMENTS

This research is funded by the ESPRIT EP2152 (VIEWS) project. The author wishes to thank Chris Brown for inspiring the exploitation of VAHMM for modelling selective attentions in visual observation, Ray Rimey for discussions on the implementation of VAHMM training, Yiannis Aloimonos for useful remarks on purposive vision systems, Graham Hill for discussions on the integration of this work into VIEWS, and Hilary Buxton for constructive comments and supports throughout the project.

References

[1] Y. Aloimonos. "Purposive and Qualitative Active Vision". In *The Proceedings of DARPA Image Understanding Workshop*, Pittsburgh, Pennsylvania, U.S., September 1990.

[2] D. Ballard. "Reference Frames for Animate Vision". In *The Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, Michigan, U.S., August 1989.

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

[4] P.J. Burt. "Image Motion Analysis Made Simple and Fast, One Component at a Time". In *Proceedings of the British Machine Vision Conference*, Glasgow, Scotland, September 1991.

[5] J.J. Clark and N.J. Ferrier. "Modal Control of an Attentive Vision System". In *The Proceedings of the Second International Conference on Computer Vision*, Tampa, Florida, U.S., December 1988.

[6] L.S. Dreschler and H-H. Nagel. "On the Selection of Critical Points and Local Curvature Extrema of Region Boundaries for Interframe Matching". In *International Joint Conference on Artificial Intelligence*, pages 542–544, 1981.

[7] S.G. Gong and J.M. Brady. "Parallel Computation of Optic Flow". In *European Conference on Computer Vision*, pages 124–133, Antibes, France, April 1990.

[8] I.E. Gordon. *Theories of Visual Perception*. John Wiley & Sons, Chichester, England, 1989.

[9] E.C. Hildreth. *The Measurement of Visual Motion*. MIT Press, Cambridge, Massachusetts, 1984.

[10] D.C. Hogg. "Finding a Known Object Using a Generate and Test Strategy". In I. Page, editor, *Parallel Architectures and Computer Vision*, pages 119–133. Oxford University Press, 1988.

[11] D.G. Lowe. "Stabilized Solution for 3-D Model Parameters". In *European Conference on Computer Vision*, pages 408–412, 1990.

[12] R. Marslin, G.D. Sullivan, and K.D. Baker. "Kalman Filters in Constrained Model Based Tracking". In *Proceedings of the British Machine Vision Conference*, Glasgow, Scotland, September 1991.

[13] R.J. McEliece, R.B. Ash, and C. Ash. *Introduction to Discrete Mathematics*. McGraw-Hill, 1989.

[14] J. Pearl. *Probabilistic Reasoning in Intelligent Systems, Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, California, 1988.

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

[16] R.D. Rimey and C.M. Brown. "Selective Attention as Sequential Behavior: Modeling Eye Movements with an Augmented Hidden Markov Model". In *The Proceedings of DARPA Image Understanding Workshop*, Pittsburgh, Pennsylvania, U.S., September 1990.

[17] A. Sloman. What are the Purposes of Vision? Technical report, University of Sussex, Brighton, UK, 1986. Cognitive Studies Research Papers CSRP0-66.

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

[19] J.K. Tsotsos. "A Complexity Level Analysis of Immediate Vision". *International Journal of Computer Vision*, 1(4), 1987.

[20] S.D. Whitehead and D.H. Ballard. Active Perception and Reinforcement Learning. Technical report, Computer Science Dept., University of Rochester, U.S.A., 1990.

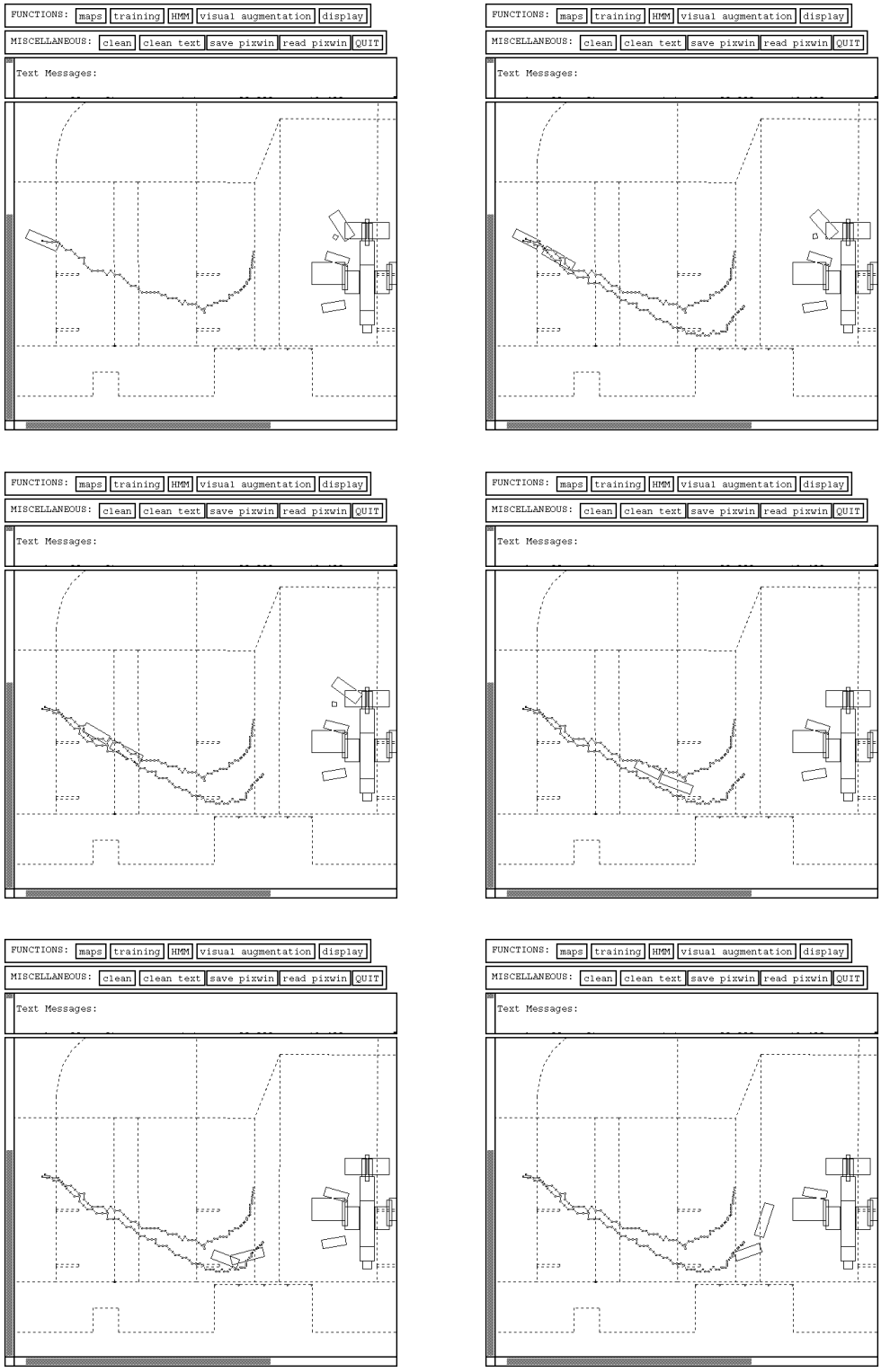

Figure 5: Once a fuel tanker appears, a visual insensitive, therefore "blind", expectation of future moving pattern of the vehicle is produced by the trained fuel tanker HMM. After 9 frames, a similar "blind" expectation is caused by the appearance of a fuel trailer. The next 4 ground plane diagrams show the overlapped initial predictions with the actual movement of the vehicles at every 20 frames interval.

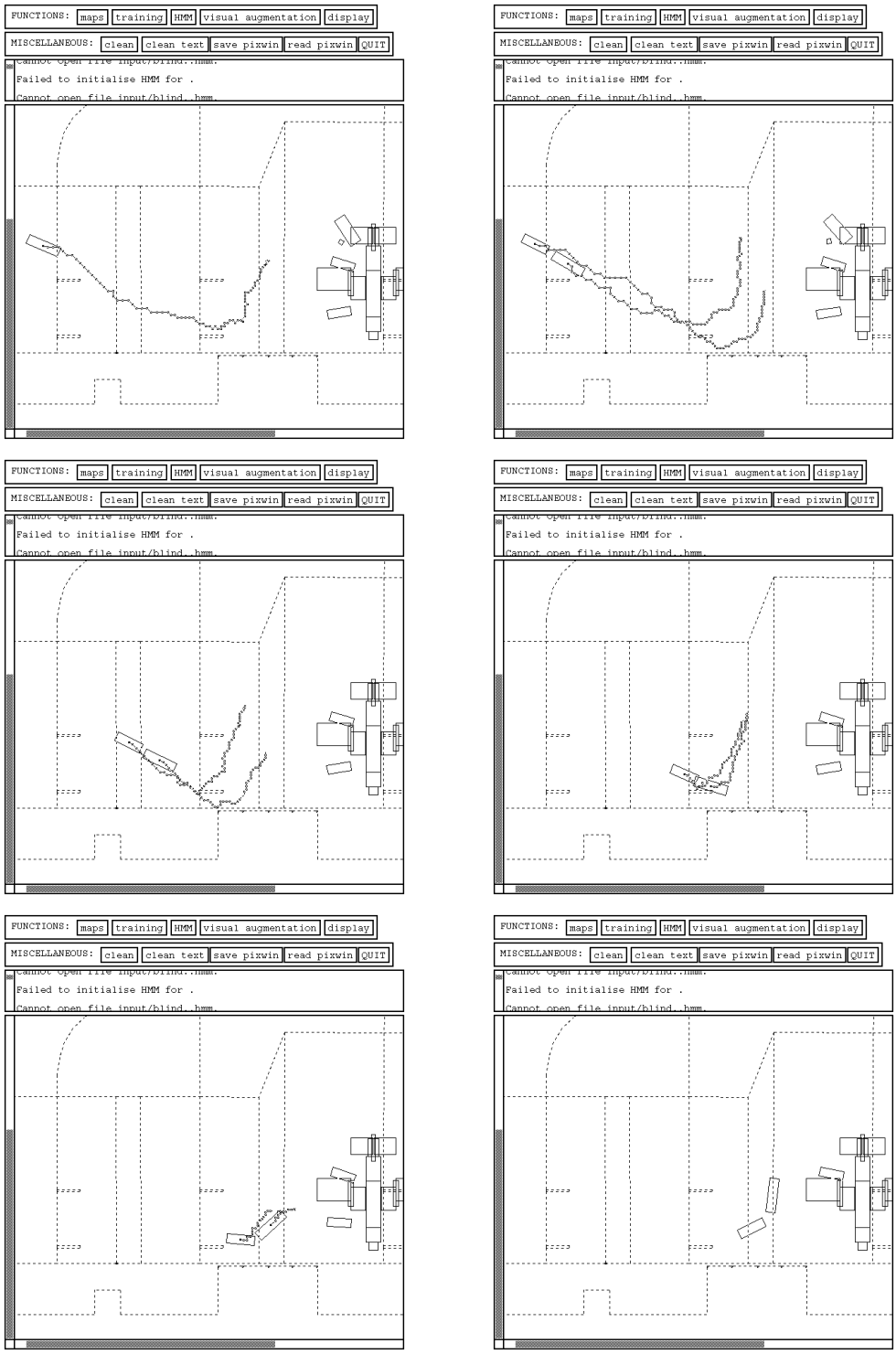

Figure 6: As the same scenario as in figure 5 but with visual augmentation at each frame. The gains from the visual augmentation are $s_{mg} = 0.1$, $s_{dg} = 0.1$, $r_{mg} = 0.4$ and $r_{dg} = 0.4$. The state transition gains should always be much smaller compared with the observation symbol gains because 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 1, an assumption was made that visual evidence collected at each frame was the result from a tracking process that has taken last frame's expectation into account. Therefore, the confidence in the visual evidence under the current circumstance is low and all the gains for the augmentation have been set low.