Spatio-temporal Discovery: Appearance + Behavior = Agent

Prithwijit Guha[1], Amitabha Mukerjee[2], and K.S. Venkatesh[1]

[1] Department of Electrical Engineering,
Indian Institute of Technology, Kanpur,
Kanpur - 208016, Uttar Pradesh
{pguha, venkats}@iitk.ac.in
[2] Department of Computer Science & Engineering,
Indian Institute of Technology, Kanpur,
Kanpur - 208016, Uttar Pradesh
amit@cse.iitk.ac.in

Abstract. Experiments in infant category formation indicate a strong role for temporal continuity and change in perceptual categorization. Computational approaches to model discovery in vision have traditionally focused on static images, with appearance features such as shape playing an important role. In this work, we consider integrating agent behaviors with shape for the purpose of agent discovery. Improved algorithms for video segmentation and tracking under occlusion enable us to construct models that characterize agents in terms of motion and interaction with other objects. We present a preliminary approach for discovering agents based on a combination of appearance and motion histories. Using uncalibrated camera images, we characterize objects discovered in the scene by their shape and motion attributes, and cluster these using agglomerative hierarchical clustering. Even with very simple feature sets, initial results suggest that the approach forms reasonable clusters for diverse categories such as people, and for very distinct clusters (animals), and performs above average on other classes.

1 Introduction

Our concepts carve nature into chunks that form more compact and efficient representations of the world. It is possible that much of our early categories are learned from a single system of knowledge, based ultimately on perception [1]. If so, how does one go about discovering these categories from the *passing show*? This is clearly one of the central problems of perception, and we attempt to approach this problem from a computational standpoint.

Temporal continuity and change plays a strong role in category formation. By three months, infants begin to pay attention to coherently moving blobs (Spelke objects [2]), and by six months they are sensitive to the spatio-temporal dynamics of occlusion [3]. Indeed, in significant respects, the behavior of objects may constitute a more important hallmark of their categorization (e.g. animacy) than appearance alone [4].

Computational models of object categorization and object recognition, on the other hand, have focused traditionally on clustering based on appearance attributes in static images [5]. While appearance attributes may be prior (e.g. faces), category formation seems to be strongly tied to dynamic scenes.

P. Kalra and S. Peleg (Eds.): ICVGIP 2006, LNCS 4338, pp. 516–527, 2006.

In this work, we consider dynamic image sequences and use improved algorithms for video segmentation and tracking under occlusion to construct coherent motion histories and occlusion relations simulating these cognitive aspects of infant category formation. The aspects of the scene that characterize agents may involve motion, shape and their interactions with other objects, and we present an attempt to form clusters based on the first two elements, and to form a set of features for the third. Thus, we obtain shape and motion characteristic for each agent instance and use an average-link hierarchical clustering algorithm to cluster agents in this combined-feature scenario. Comparing our results with those obtained based on appearance alone shows significant improvements in recognizing certain heterogeneous groups such as People. Thus, through this work, we present an initial approach for discovering agents based on a combination of appearance and motion histories.

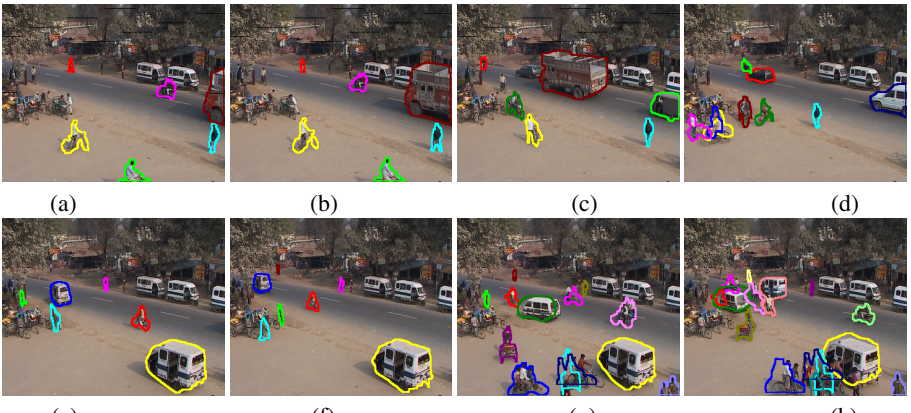


Fig. 1. Image sequence showing multi-agent activity in a traffic surveillance scenario. Vehicles, people, and animals are tracked across partial and complete occlusions and identified as agent instances. Frames (a)-(h) shows a crowd of people embarking a static tempo (marked as an erstwhile moving object) in the foreground, while a truck, motorcycles, rickshaws, bicycles and pedestrians are seen interacting along the main road.

1.1 Dynamic Image Characterization

We use an image sequence (acquired with a static camera) involving simultaneous interactions of tens of agents in a traffic scene, interacting with and occluding each other (figure 1). Agents are identified as connected blobs that are capable of motion, including those that are currently at rest. Agents constitute 3-manifolds in the $\{\text{image}\times\text{time}\}$ space, which is reduced to a set of features covering appearance and behavior.

Behavior of an agent can involve many aspects. Here we distinguish solitary behaviors (motions) from multi-agent behaviors (interactions). Based on the assumption that interaction involving proximity in $3D$ space may result in visual occlusion, we use occlusion, as one of the indicators of interaction. These three sets of features, then, are central to agent recognition:

- Appearance: attributes include shape, which is captured via the summary attributes of area, aspect ratio, and dispersedness (Perimeter^2 / Area).
- Motion : attributes include speed, direction change sequences, velocity change, as well as the quadratic-splined curve of the trajectory.
- Interactions: These are computed based on occlusion primitives (sub-section 2.3) such as isolation, crowding, fragmentation and disappearance.

Next we highlight our approach to multi-agent tracking in crowded scenarios. Since these algorithms have already been presented elsewhere [6] only a minimal summary is presented below.

2 Multiple Agent Tracking

Connected regions in the foreground (sub-section 2.1) that are seen to persist coherently over time constitute agents. Agents are tracked across multiple frames, and partial or complete occlusions are handled. The part of the image corresponding to an agent (called its support region) is tracked across frames by associating the predicted agent regions to foreground blobs obtained from figure-ground segmentation (2.1).

Several agents may correspond to the same blob (crowding), or a single agent may map to several blobs (fragmentation), etc; and the nature of this mapping is also stored in terms of several occlusion primitives (sub-section 2.3). The agents are further localized by an iterative centroid update algorithm, where their appearance (position and color) features are used to re-estimate the agent centroids (2.2). For the purpose of prediction, agent-blob associations are re-computed and agent models updated only for those agents which are unoccluded by others.

Agents which have been successfully tracked at the t^{th} instant comprise the *active* set of agents $\mathcal{A}_{active}(t)$, whereas agents which have disappeared (occluded by others, etc.) constitute the set $\mathcal{A}_{disapp}(t)$ - which are matched with new agents when they arise in the scene. The system initializes itself with empty sets and agent instances are added (removed) as they appear (disappear) in the field of view.

2.1 Foreground Extraction

Agents are identified as foreground regions based on one of two kinds of evidence: first, as regions of change with respect to a learned background model; and second, as regions exhibiting motion. The background model is learned as a pixel-wise mixture of Gaussians [7] only for those pixels which exhibit no image motion. Inter-frame motion information [8] along with the higher level multi-agent tracking feedback (sub-section 2.2) is used further to disambiguate objects that come to a stop or objects that suddenly start moving [6]. The detected foreground pixels are further subjected to neighborhood-voting based corrections followed by connected component analysis to obtain a set of disjoint foreground blobs $\mathcal{F}(t) = \{f_i(t) : i = 1, \ldots, n_t\}$.

2.2 Agent Localization

Tracking multiple agents involve the use of their appearance models along with the trajectory information. The appearance of the j^{th} agent having region of support consisting

of $n_j(t)$ pixels at the t^{th} instant is represented by an appearance model $\mathbf{a}_j(t) = \{\alpha_{jk}(t) : k = 1, \ldots, n_j(t)\}$, where $\alpha_{jk}(t) = [r_{jk}(t), g_{jk}(t), b_{jk}(t), \tilde{x}_{jk}(t), \tilde{y}_{jk}(t)]^T$ characterizes the k^{th} component pixel of the j^{th} agent. Here, the appearance is a collection of pixel positions $(\tilde{x}_{jk}(t), \tilde{y}_{jk}(t))$ relative to the centroid $\mathbf{c}_j(t)$ of the agent pixel set and the corresponding colors $(r_{jk}(t), g_{jk}(t), b_{jk}(t))$ in the RGB color space.

The agent-blob association is performed among the members of the active set $\mathcal{A}_{active}(t)$ and the set of foreground blobs $\mathcal{F}(t)$. If at least one pixel of $\mathbf{a}_j(t)$ overlaps $f_i(t) : |\{(\tilde{x}_{jk}(t), \tilde{y}_{jk}(t)) + \mathbf{c}*_j(t)\} \cap f_i(t)| \neq 0$, where $\mathbf{c}*_j(t)$ is the predicted agent centroid obtained at the t^{th} instant from the motion history.

We consider the general case where the agents $\{A_p(t) : p = 1, \ldots, P_t\}$ are associated with the foreground blobs $\{f_q(t) : q = 1, \ldots, Q_t\}$. Let, $\mathbf{u}_l(t)$ be a foreground pixel characterized by color and position as $\mathbf{u}_l(t) = [r_l(t), g_l(t), b_l(t), x_l(t), y_l(t)]^T$. We consider $\mathbf{u}_l(t)$ to be the best match of the k^* pixel of the j^* associated agent, if the following are satisfied.

$$\tilde{\mathbf{u}}_{lj}(t) = [r_l(t), g_l(t), b_l(t), (x_l(t), y_l(t)) - \mathbf{c}_j^{(s)}(t)]^T \tag{1}$$

$$\lambda_{jkl}^{(s)}(t) = exp(-\|\alpha_{jk}(t) - \tilde{\mathbf{u}}_{lj}(t)\|_{\Sigma}) \tag{2}$$

$$v_l^{(s)}(t) = (j^*, k^*) = argmax_{j,k}\lambda_{jkl}^{(s)}(t) \tag{3}$$

where Σ is a weighing matrix and $\mathbf{c}_j^{(s)}(t)$ is the iteratively re-estimated centroid of the support of the j^{th} agent in the s^{th} iteration for the t^{th} frame. The best match agent-pixel index two tuple for $\mathbf{u}_l(t)$ is given by $v_l^{(s)}(t) = (j^*, k^*)$, which indicates the visibility of the k^* pixel of the j^* agent as $\mathbf{u}_l(t)$. Thus, with respect to $\mathbf{u}_l(t)$, the centroid of $A_{j^*}(t)$ can be expected to be at $\hat{\mathbf{c}}_{j^*k^*l}^{(s)}(t) = (x_l(t), y_l(t)) - (\tilde{x}_{j^*k^*}(t), \tilde{y}_{j^*k^*}(t))$ with a certain weighted belief $\lambda_{jkl}^{(s)}(t)$. We re-estimate the agent centroid in the $(s+1)^{th}$ iteration as the weighted average of such expected centroid positions derived in s^{th} iteration as shown in equation 5.

$$V_j^{(s)}(t) = \{l, k : v_l^{(s)}(t) = (j, k)\} \tag{4}$$

$$\mathbf{c}_j^{(s+1)}(t) = \frac{\sum_{l,k \in V_j^{(s)}(t)} \hat{\mathbf{c}}_{jkl}^{(s)}(t)\lambda_{jkl}^{(s)}(t)}{\sum_{l,k \in V_j^{(s)}(t)} \lambda_{jkl}^{(s)}(t)} \tag{5}$$

The agent localization iterations are initialized with $\mathbf{c}_j^{(0)}(t) = \mathbf{c}*_j(t)$ and are terminated, when $max_j\|\mathbf{c}_j^{(s+1)}(t) - \mathbf{c}_j^{(s)}(t)\| \leq \epsilon_c$ is satisfied. The experiments are performed on 5000 frames of a traffic surveillance video, acquired with a static camera under almost constant ambient illumination conditions. We have used a diagonal weighing matrix (equation 2) of $\Sigma = [0.25, 0.25, 0.25, 1, 1]$ and obtained a tracking accuracy of 61%. The results of multiple agent tracking on the traffic video are shown in figure 1.

2.3 Interactions: Occlusion Primitives

Interactions between objects in $3D$ cannot be dealt with, but one may assume that objects in close proximity are likely to occlude one another (given a supra-horizon view).

The nature of this occlusion and its transitions constitute a partial signature of the interaction. Four types of occlusion situations are distinguished:

- Isolation ($\mathcal{O}(I)$): Single agent associated to single foreground blob (No occlusion).
- Crowding ($\mathcal{O}(C)$): More than one agents are associated to a single foreground blob (Occluding or occluded by other agents).
- Fragmentation ($\mathcal{O}(P)$): Agent appears as fragmented, being associated to multiple foreground blobs (partial occlusions).
- Disappearance ($\mathbf{O}(D)$): Agent unassociated to any foreground blob (Complete occlusion by background objects).

In *crowding / fragmentation* situations, only the agent trajectory is updated. Appearance models are updated only for those agents which are unoccluded by other agents/background objects. *Disappeared* agents are moved from $\mathcal{A}_{active}$ to $\mathcal{A}_{disapp}$. Foreground blobs (or fragments) unassociated to any agent in $\mathcal{A}_{active}$, are compared against those in $\mathcal{A}_{disapp}$ - if a match is not found, an entrance of a new agent is declared, otherwise the matched agent is reinstated in $\mathcal{A}_{active}$. Clearly, since our recognition is $2D$, many agents which re-emerge after near-complete occlusion may not be recognized - and indeed this is the case in approximately half of such situations.

3 Agent Characterization

An agent instance $A_j(t_s(j), t_e(j))$ is a space-time manifold characterized by the time indexed set of appearances (a collection of position (XY) and corresponding color (RGB) vectors) and the centroid-trajectory $\{\mathbf{c}_j(t) : t = t_s(j), \ldots, t_e(j)\}$ during its scene presence $[t_s(j), t_e(j))$. This agent model encodes both the appearance and motion and constitutes part of the cognitive percept of the agent; the other aspects being its interactions with other objects.

This work focuses on unsupervised agent categorization based on low level features derived from the shape and the motion. As shape features, we consider the area (number of pixels in the support region), aspect ratio (vertical to horizontal length ratio of the minimum bounding box of the agent) and dispersedness (ratio of perimeter squared to area). Collins et al. [9] have successfully classified people and vehicles using this simple set of features, albeit in a supervised learning framework. The motion features include the set of speeds, directions of motion and the form (e.g. linear versus quadratic) of the trajectory in the image space. Such features can assist us in handling queries such as "*List all high speed vehicles*", "*List all agents moving in a straight path from left to right*" or "*Group agents of similar size*" etc. The distributions of the shape and motion features of all the agents are shown in figure 2. Categorizing agents with respect to their low level features are described in further details in sub-section 3.1.

Motion and interactions together constitute the behavior model of an agent. Interactions in the real world often leave their imprints in terms of image space occlusions. Thus, the interactions with other objects in the scene can be represented by the occlusion primitive transition sequences of the concerned agents. This might lead to behavioral descriptions like "*The agent that walked across the tree*".

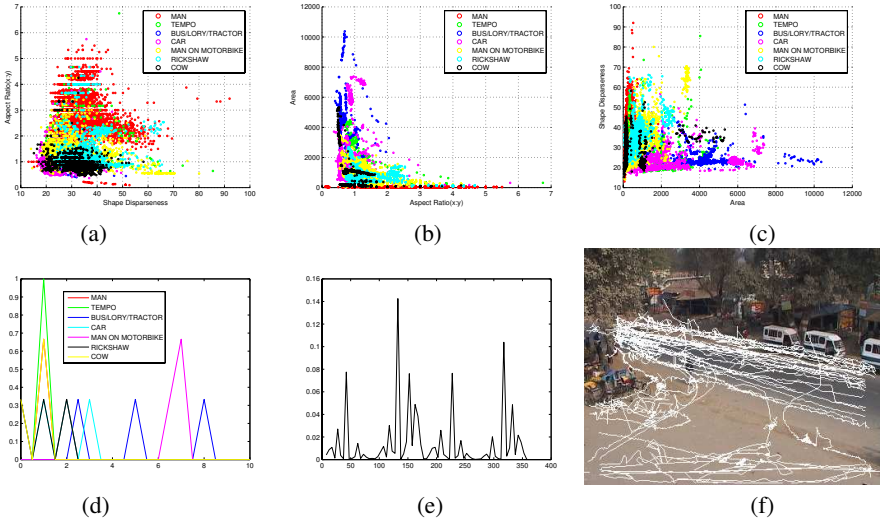


Fig. 2. Low level shape features projected on the (a) aspect ratio - dispersedness plane, (b) area - aspect ratio plane and (c) dispersedness - area plane; (d) Agent speed distribution in image plane; (e) Distribution of directions of motion in image plane; (f) Agent trajectories obtained till the first 2500 frames

3.1 Agent Categorization

The scene presence of the agents vary leading to the formation of variable length shape and motion feature sets. Here, the agent A_r is represented as a set of certain features q as $A_r^{(q)} = \{\mathbf{q}_{rj} : j = 1, \ldots, n_r^{(q)}\}$. Such forms of agent characterizations can't be efficiently handled by the usual approach of learning mixtures of Gaussians. Thus, we opt for "*agglomerative hierarchical clustering algorithms*" [10] which only require a definition of a distance measure between two such sets. Consider two Agents A_r and A_s characterized by their respective feature sets $A_r^{(q)}$ and $A_s^{(q)}$. We define the distance $\mathcal{D}_a(A_r^{(q)}, A_s^{(q)})$ between these two agent feature sets as,

$$\mathcal{D}_a(A_r^{(q)}, A_s^{(q)}) = \frac{1}{n_r^{(q)} n_s^{(q)}} \sum_{j=1}^{n_r^{(q)}} \sum_{k=1}^{n_s^{(q)}} \|\mathbf{q}_{rj} - \mathbf{q}_{sk}\| \tag{6}$$

We employ the *average-link clustering algorithm* as it offers a compromise between the single-link and complete-link ones and is more robust to outliers [10]. The algorithm initializes by assigning a cluster label to each of the agents. Thus, for a collection of agent feature sets $\{A_i\}_{i=1}^n$, the initial collection of clusters is given by $\{C_i = \{A_i\}\}_{i=1}^n$. A dendogram is formed in a bottom-up approach, where each iteration reduces the cluster number by one while merging two clusters, until finally, at the n^{th} iteration, all the agents are assigned to a single cluster.

Consider the k^{th} iteration, where we merge two of the $(n - k + 2)$ clusters obtained from the $(k - 1)^{th}$ step. The diameters of all possible 2-cluster mergers are computed

and the pair minimizing the same is considered for merging. We select the cluster index pair (i^*, j^*) for merging in the k^{th} iteration, if $(i^*, j^*) = argmin\mathcal{D}_c(C_i, C_j), \forall i, j = 1 \ldots (n-k+2)$ and $i \neq j$, where $\mathcal{D}_c(C_i, C_j)$ is the distance between two clusters given by,

$$\mathcal{D}_c(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{A \in C_i} \sum_{A' \in C_j} \mathcal{D}_a(A, A') \tag{7}$$

3.2 Categorizing with Shape and Motion Features

The multiple agent tracking performed on the aforementioned traffic surveillance video yielded the discovery of $n = 376$ agents. The appearances of the discovered agents are manually inspected for generating ground-truth data, from which we observe the existence of 10 distinct categories along with outliers (formed from track losses and spurious foreground detection) - $\Gamma(\text{SHAPE}) = \{$ OUTLIERS (62) PEOPLE (130), TEMPO (18), BUS (3), TRUCK (1), TRACTOR (3), CAR (18), MOTORBIKE (55), CYCLE (44), RICKSHAW (25), COW (17) $\}$.

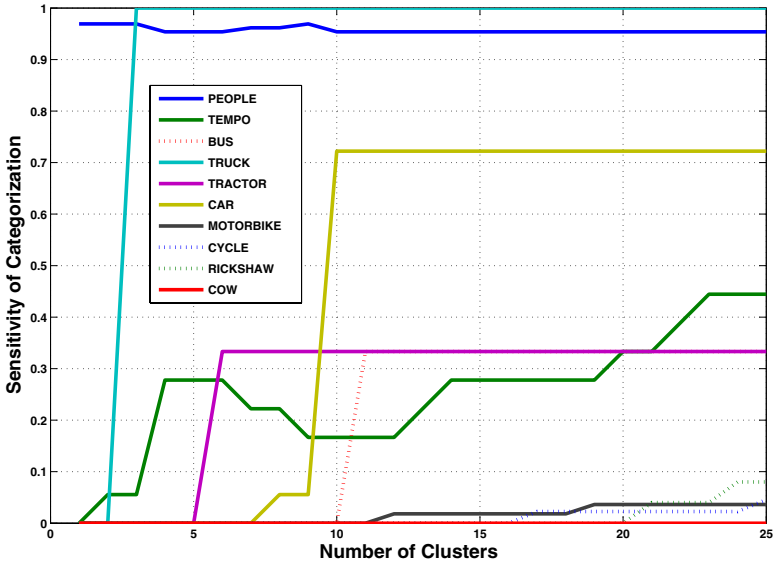
We form the collection of agent shape feature (area, aspect ratio and dispersedness) sets $\{A_i^{(shape)}\}_{i=1}^n$, which are subjected to the average link clustering algorithm. A certain cluster $C_i^{(k)}$ in the k^{th} iteration is declared to host a certain agent category Γ_j, if the agents of that class occur with the highest frequency in $C_i^{(k)}$. The sensitivity $\mathcal{S}_j(k)$ of categorizing the agent category Γ_j in the k^{th} iteration is thus defined as,

$$\mathcal{S}_j(k) = \frac{1}{|\Gamma_j|} \sum_{i=1}^{n-k+1} |C_i^{(k)}| \delta(\beta(i) - j) \tag{8}$$

Where, $|\Gamma_j|$ is the total number of instances of the j^{th} category, $\beta(i)$ denotes the category of an agent (from groundtruthed data), $|\bullet|$ determines the cardinality of a cluster and $\delta(\bullet)$ is the Kronecker Delta function. The sensitivities of clustering with up to 25 clusters for shape features and instances of the appearances of the discovered agents are shown in figure 3.

In a similar manner, we construct the agent trajectory feature sets of direction $\{A_i^{(direction)}\}_{i=1}^n$ and form $\{A_i^{(form)}\}_{i=1}^n$ which are subjected to hierarchical clustering. Manual inspection of the ground-truth data shows the existence of 6 different kinds of trajectories along with outliers (on account of track losses) - $\Gamma(\text{TRAJECTORY}) = \{$ OUTLIERS (205) LEFT TO RIGHT (77), FROM BOTTOM TURN LEFT (5), MOVE UP (4), RIGHT TO LEFT (76), U-TURN (3), MOVE TO BOTTOM (6) $\}$. Among these, there were only 163 LINEAR TRAJECTORIES. The sensitivities of detecting trajectories up to 100 clusters and the three distinctly discovered trajectories (LEFT TO RIGHT , FROM BOTTOM TURN LEFT and RIGHT TO LEFT)are shown in figure 4.

The high sensitivities (figures 3 and 4) in grouping the appearances of PEOPLE, TRUCK , CAR and the LINEAR TRAJECTORIES of agents moving LEFT TO RIGHT and RIGHT TO LEFT is indicative of the satisfactory performance of the proposed approach.



(a)

(b)

Fig. 3. Results of agent appearance discovery. Area, aspect ratio and dispersedness constitute the appearance features of an agent. However, the cardinalities of the appearance feature sets of the discovered agents vary due to their different frame presence. Here, we only have a measure of distance between two such sets and thus the average link clustering algorithm is executed on a collection of 376 (discovered) agent shape feature sets. (a) Detection rates achieved by the average link clustering algorithm computed by Cross-validating with ground-truth data. (b) Instances of appearances of discovered agents (appearances are scaled for better viewing purposes).

3.3 Behavior as Variable Length Occlusion Sequences

Models of single-agent behaviors are mainly characterized by their state space trajectories. Agent categorization in a surveillance scenario in terms of its motion features have already been discussed in sub-section 3.2. Agent-object interactions exhibit several different modes - the actions may involve actual contact (e.g. riding a bike, boarding or

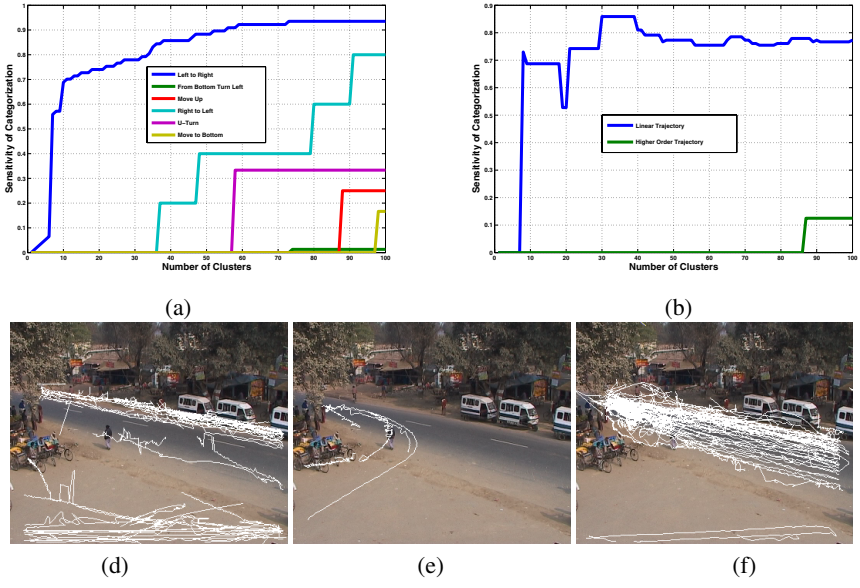


Fig. 4. Detection rates achieved by the average link clustering algorithm computed by cross-validating with ground-truth data. Sensitivity of categorization of agents by (a) *Trajectory direction* and (b) *Trajectory form* (linear versus higher order); Plotting discovered trajectories with respect to form and direction. Linear trajectories (in image plane) of agents moving (d) *Left to Right*, (e) *Right to Left* and (f) higher order trajectories of *Turning Left coming from Bottom*.

disembarking a vehicle, grouping etc.) or may involve interactions at a distance (e.g. following, chasing, overtaking, etc.). In terms of image space, actual contacts are necessarily reflected in terms of occlusions, but non-contact situations do not necessarily lead to non-overlap. Thus, we believe that occlusion sequences exhibited by an agent form visual signatures for the underlying interactions. More so, we identify that the occlusion state transition sequences form a more significant interaction description than the occlusion state sequences themselves. In this work, we aim to discover the interactions arising out of agents moving in complex environments undergoing both static and dynamic occlusions with background objects and other agents respectively.

A number of methodologies employing hidden Markov models, time-delay neural networks, recurrent networks etc. have been proposed for modeling and recognition of action/interaction sequences in a supervised learning framework. On the other hand, unsupervised learning of activity patterns have also been proposed by trajectory clustering [11] or variable length Markov model learning [12]. A good overview of such techniques can be found in [13].

Supervised activity modeling techniques are mostly task oriented and hence fail to capture the corpus of events from the time-series data provided to the system. Unsupervised data mining algorithms, on the other hand, discover the modes of spatio-temporal patterns thereby leading to the identification of a larger class of events. The use of VLMMs in the domain of activity analysis was introduced for automatic modeling of

the actions in exercise sequences [14] and interactions like handshaking [12] or overtaking of vehicles [15] in a traffic scenario. These approaches propose to perform a vector quantization over the agent feature and motion space to generate temporally indexed agent-state sequences from video data. These sequences are parsed further to learn VLMMs leading to the discovery of behavioral models of varying temporal durations.

Incremental Transition Sequence Learning. In this work, we employ "*Incremental Transition Sequence Learning*" to capture the variable length sequences of occlusion primitives which describe different behavior patterns. The atomic event primitives (here, the occlusion primitives) ϵ constitute the set $\mathcal{E}$. Our approach to mining in this space involves the construction of an *activity tree* $\mathcal{T}_\alpha$ whose branches represent variable length event primitive sequences.

An empty (first in first out) buffer β_j (of length L, the maximum sequence length) and the null activity tree $\mathcal{T}_\alpha(j)$ (containing only the root node ρ_j) are initialized at the very first appearance of every j^{th} agent $\mathcal{A}_j$. Each node of $\mathcal{T}_\alpha(j)$ is a two tuple $\mathcal{T}_n \equiv (\epsilon, \pi)$ containing the primitive $\epsilon \in \mathcal{E}$ and a real number $\pi \in (0, 1]$ signifying the probability of occurrence of the path $\{\rho_j, \ldots, \mathcal{T}_n\}$ among the set of all possible paths of the same length.

Let, $\epsilon(j,t)$ be the event primitive observed for $\mathcal{A}_j$ at time t. If there is a transition in this event primitive, i.e. if $\epsilon(j,t) \neq \epsilon(j,t-1)$, then $\epsilon(j,t)$ is pushed to β_j. Let the set of l-length paths (originating from ρ_j) of $\mathcal{T}_\alpha(j,t)$, be $B^{(l)}(j,t) = \{\alpha_u^{(l)}(j,t)\}_{u=1}^{b_l}$, where b_l is the number of l-length branches in the tree. More so, if the sequence $\{\beta_j[l-k](t)\}_{k=1}^{l}$ signify the b^{th} path of $B^{(l)}(j,t)$, then the probabilities $\{\pi_u^{(l)}(j,t)\}_{u=1}^{b_l}$ of the nodes of $\mathcal{T}_\alpha(j,t)$ at the l^{th} depth are updated as,

$$\pi_u^{(l)}(j,t) = (1-\eta_l(t))\pi_u^{(l)}(j,t-1) + \eta_l(t)\delta(u-b) \tag{9}$$

Where, $\eta_l(t)$ is the rate of learning l-length sequences at the t^{th} instant and δ is the Kronecker delta function. However, in the current implementation a fixed learning rate η is employed such that $\eta_l(t) = max\left(\frac{1}{t}, \eta\right) \forall l$.

Occurrence of a new event primitive results in the formation of newer variable length sequences in the buffer. Thus, new nodes signifying this event primtive are added at different levels of the tree thereby growing newer branches. Each new node is initialized with an initial probability of $\eta_l(t)$, whereas the older node probabilities in the same levels are penalized by multiplying with a factor of $(1-\eta_l(t))$. This ensures the self-normalizing nature of node probability updates (as in equation 9) such that they add up to unity at each depth.

3.4 Occlusion Interaction: Learning from O-Transitions

Consider a short video sequence where a person walks across a tree from left to right in the image space from which we sample 18 frames to illustrate the process of agent-background object interaction discovery. Key frames from this sequence are shown in figure 5(a)-(e). Incremental transition sequence learning is performed with a maximum depth of $L = 10$ and a learning rate η inversely proportional to the frame number. The growth of the activity tree is shown in figure 5(f).

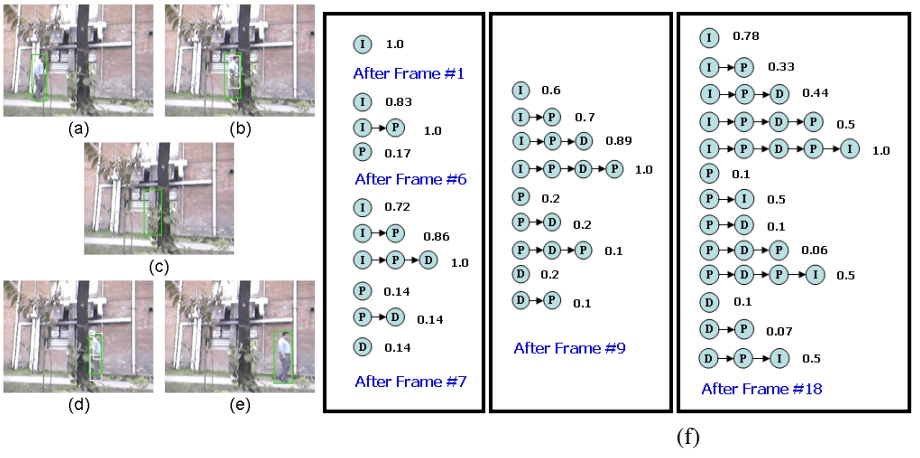


Fig. 5. Example video sequence: Man walks left to right behind a tree. Frames and Agent states: (a) 1-5: *isolated*; (b) 6: *partially occluded*, (c) 7-8: *disappeared*, (d) 9: *partially occluded* (e) 10-18: *isolated*. (f) Learning **Activity Tree**. The left-most nodes are just below the root of the growing tree. Results of incremental transition sequence learning are shown after frames 1, 6, 7, 9 and 18. Branches encode different variable length event sequences along with relative frequencies. Thus, in column 2 (after Frame 9), the sequence $\{(I \rightarrow P \rightarrow D), 0.89\}$ corresponds to the event primitive sequence $\{\mathcal{O}(I) \rightarrow \mathcal{O}(P) \rightarrow \mathcal{O}(D)\}$; i.e. the event sequence "coming from the left and getting hidden" occurs with relative frequency 89% among observed 3-length sequences.

Semantic labels can be assigned to the sequences in the occlusion-primitive space to denote different activities, and subsequences may constitute sub-activities. For example, consider the longest path $\{\mathcal{O}(I) \rightarrow \mathcal{O}(P) \rightarrow \mathcal{O}(D) \rightarrow \mathcal{O}(P) \rightarrow \mathcal{O}(I)\}$ learned in the activity tree from the aforementioned video that correspond to the activity of **walking across a tree from left to right**. Subsequences of this path viz. $\{\mathcal{O}(I) \rightarrow \mathcal{O}(P) \rightarrow \mathcal{O}(D)\}$ and $\{\mathcal{O}(D) \rightarrow \mathcal{O}(P) \rightarrow \mathcal{O}(I)\}$ also correspond to the visually significant events of **going to hide from left to right** and **reappearing and moving to the right**.

4 Conclusion

In this work we have attempted to capture some cognitive notions of perceptual category learning and attempted to devise computational analogs of this process. While our results in terms of category discovery are clearly preliminary, we believe that the high correlations obtained in terms of sensitivity matches (figures 3 and 4) do indeed provide some justification for such an approach, especially given that the system operates in completely unsupervised mode, without any information about the complex scene that is being observed.

Clearly, the results are indicative, and much work remains. In particular, characterizing the interactions between objects is a very rich area, of which the occlusion models used here only scratch the surface. In cognitive terms, the behavior of agents can be described compactly in terms of landmarks - and currently we are exploring the emergence of scene landmarks with which to characterize such interaction.

References

1. Quinn, P., Eimas, P.: The emergence of category representations during infancy: Are separate perceptual and conceptual processes required? Journal of Cognition and Development **1** (2000) 55–61
2. Spelke, E.S.: Principles of object perception. Cognitive Science **14** (1990) 29–56
3. Gredebaeck, G., von Hofsten, C.: Infants evolving representations of object motion during occlusion: A longitudinal study of 6- to 12-month-old infants. Infancy **6** (2004) 165–184
4. Mandler, J.M.: Foundations of Mind. Oxford University Press, New York (2004)
5. Mareschal, D., French, R.M., Quinn, P.: A connectionist account of asymmetric category learning in early infancy. Developmental Psychology **36** (2000) 635–645
6. Blind1: Details withheld. In: for Blind Review. (2006)
7. Zivkovic, Z.: Improved adaptive gaussian mixture model for background subtraction. In: Proceedings of the 17th International Conference on Pattern Recognition. Volume 2. (2004) 28–31
8. Proesmans, M., Gool, L.V., Pauwels, E., Osterlinck, A.: Determination of optical flow and its discontinuities using non-linear diffusion. In: The 3rd Eurpoean Conference on Computer Vision. Volume 2. (1994) 295–304
9. Collins, Lipton, Kanade, Fujiyoshi, Duggins, Tsin, Tolliver, Enomoto, Hasegawa: A system for video surveillance and monitoring: Vsam final report. Technical Report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University (2000)
10. Duda, R., Hart, P., Stork, D.: Pattern Recognition. 2nd Edition, John Wiley and Sons (2003)
11. Johnson, N., Hogg, D.: Learning the distribution of object trajectories for event recognition. In: Proceedings of the 6th British conference on Machine vision (Vol. 2), BMVA Press (1995) 583–592
12. Galata, A., Johnson, N., Hogg, D.: Learning variable-length markov models of behavior. Computer Vision and Image Understanding **81** (2001) 398–413
13. Buxton, H.: Learning and understanding dynamic scene activity: a review. Image and Vision Computing **21** (2003) 125–136
14. Johnson, N., Galata, A., Hogg, D.: The acquisition and use of interaction behavior models. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society (1998) 866–871
15. Galata, A., Cohn, A.G., Magee, D., Hogg, D.: Modeling interaction using learnt qualitative spatio-temporal relations and variable length markov models. In van Harmelen, F., ed.: Proceedings of European Conference on Artificial Intelligence. (2002) 741–745