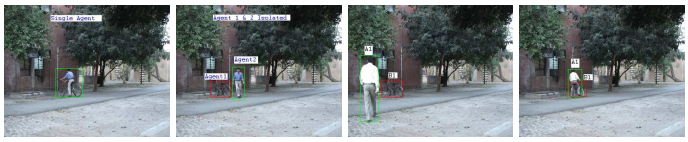
Activity Discovery from Occlusion Primitives

Abstract

Complex multi-agent interactions result in occlusion sequences which are a visual signature for the event. In this work, multi-agent interactions are tracked using a set of qualitative occlusion primitives derived based on the Persistence Hypothesis (objects continue to exist even when hidden from view). Variable length temporal sequences of occlusion primitives are shown to be well-correlated with many classes of semantically significant events. In surveillance applications, determining occlusion primitives is based on foreground blob tracking, and requires no prior knowledge of the domain or camera calibration. New foreground blobs are identified as putative agents which may undergo occlusions, split into multiple agents, merge back again, etc. Temporally significant sequences are identified through temporal sequence mining, and these bear high correlation with semantic categories (e.g. disembarking from a vehicle involves a series of splits). Thus semantically significant event categroies can be recognized without assuming camera calibration or any environment/agent/action model priors.

1 Introduction

Systems that detect events in multi-agent interactions often treat occlusions as a problem. On the contrary, we believe that temporal sequences of occlusion phenomena constitute a qualitative signature of the underlying event. Unlike quantitative approaches using supervised priors for object / behaviour recognition, (e.g. [5, 8, 10]), occlusion signatures are fundamental to activity in a manner that is independent of imaging and does not require any priors for either agents or events, and may have cognitive correlates with developmental processes in early vision, as called for in [6]. We construct variable length Markov models (VLMM) [4, 3] for action discovery from temporal sequences of occlusion primitives, and generate signatures for a wide class of actions. For example, a group of people hugging each other, a person coming on a bicycle, getting off and going into a building, a crowd of people coming out of a bus (Figure 1), are all events which have stable signatures in terms of occlusion primitives (henceforth o-primitive).



(m) (n) (o) (p)

Figure 1. A putative agent blob enters from the right, and a person (A2) emerges (splits) out, leaving A1 (bicycle) in a state of stasis. A2 exits the scene and A1 is layered onto the image background. A new agent A3 enters from left, and merges with A1, and together the merged blob exits from the right. Note that the semantically significant aspects of this event maps to a sequence of o-primitives which are independent of the shapes of the objects, the viewpoint, etc.

2 Occlusion Primitives

The algorithm works by tracking the foreground blobs and labeling several characteristic occlusion behaviours accoridng to the **Persistence Hypothesis**: Objects continue to exist even when hidden from view. This results in six possible states for an agent with respect to other objects/agents: *isolation, partial occlusion* (several foreground objects in one agent region), *crowding* (several agents in the same foreground region), *disappeared* (full occlusion or tracking lost). Additionally, two special primitives - *entrance* and *exit* refer to occlusion by the viewing frame itself. Figure 2 shows examples of different parts of a person visible from behind a tree (partial occlusion), or several people merged together (crowding).

Algorithm Overview: After learning a background model, foreground blobs (putative agents) are tracked and o-primitives associated with agents are noted. Frequently occurring temporal chains of variable length are identified as events of possible semantic interest.

2.1 Identifying Occlusions

The background model is learned based on a multi-scale co-linearity statistic [7]. In the current demonstration version, the first T frames $\{\Omega_t\}_{t=1}^{T}$ (in normalized RGB color space) are assumed to be unintruded by any agent(s). The second order statistics of the background features are employed to classify the set of foreground pixels in a new image Ω_t ($t > T$). The detected foreground pixels are subjected to further post-processing stages (shadow suppression, morphological corrections and connected component analysis) to obtain the set of disjoint foreground blobs at the t^{th} instant as $\mathcal{F}_t = \{F_i(t)\}_{i=1}^{n_t}$.

(a) (b)

Figure 2. Cases of occlusions. (a) Partial occlusion: agent occluded by tree is visibile as three fragmented blobs; in this state, the agent is recognized, but its visual characteristics are not updated. (b) Crowding: multiple agents merge to form a single blob. Each agent in the blob must be partially visible (otherwise it is flagged as *disappear*).

The agent-blob association is performed over an *active* set $\mathcal{S}_A(t) = \{\mathcal{A}_j(t)\}_{j=1}^{m_t}$ containing agents tracked till the t^{th} instant and also a *putative* set $\mathcal{S}_P(t-1)$ of agents of which have disappeared within the viewing window. The system initializes itself with empty sets and the agents are added (removed) as they (dis)appear in (from) the field of view. The j^{th} agent in $\mathcal{S}_A(t)$ is characterized by its occupied pixel set $a_j(t)$, weighted color distribution $h_j(t)$ and the order-τ trajectory of the center $c_j(t)$ of the minimum bounding rectangle of $a_j(t)$. Mean-shift iterations [2] initialized with the motion predicted position from the trajectory $\{c_j(t-t')\}_{t'=1}^{\tau}$ are used to localize $\mathcal{A}_j(t)$ in the t^{th} frame. To associate the agent $\mathcal{A}_j(t)$ with the foreground blob $F_i(t)$, we construct the thresholded *localization confidence matrix* $\Theta_{AF}(t)$ and the *attribution confidence matrix* $\Psi_{FA}(t)$. These confidences are computed by a fractional overlap measure $\gamma(\omega_1, \omega_2) = \frac{|\omega_1 \cap \omega_2|}{|\omega_1|}$ signifying the fraction of the region ω_1 overlapped with ω_2.

$$\Theta_{AF}[j,i](t) = \begin{cases} 1; & \gamma(A_j(t), F_i(t)) \geq \eta_A \\ 0; & \text{Otherwise} \end{cases} \quad (1)$$

$$\Psi_{FA}[i,j](t) = \begin{cases} 1; & \gamma(F_i(t), A_j(t)) \geq \eta_F \\ 0; & \text{Otherwise} \end{cases} \quad (2)$$

The number of foreground regions attributed to the j^{th} agent ($\Theta_A[j](t) = \sum_{i=1}^{n_t} \Theta_{AF}[j,i](t)$ and $\Psi_A[j](t) = \sum_{i=1}^{n_t} \Psi_{FA}[i,j](t)$) and agents localized in $F_i(t)$ ($\Theta_F[i](t) = \sum_{j=1}^{m_{t-1}} \Theta_{AF}[j,i](t)$ and $\Psi_F[i](t) = \sum_{j=1}^{m_{t-1}} \Psi_{FA}[i,j](t)$) are further computed from these matrices. The j^{th} agent in $\mathcal{S}_A(t-1)$ is **isolated** (unoccluded) ($\epsilon(I)[j,t]$), if the localization confidence is significantly high and the associated foregorund blob is not overlapped with other agents. However, when both localization and attribution confidences fall below η_A and η_F, the agent has (**disappeared**) ($\epsilon(D)[j,t]$). In case of **partial occlusions** ($\epsilon(P)[j,t]$), the attribution confidence of one or more foreground blobs to the j^{th} agent remains high, although the localization confidence falls significantly. On the other hand, while in a **crowd** ($\epsilon(C)[j,t]$), the localization confidence of the j^{th} agent in the crowded blob (overlapped with more than one agent) remains high although the attribution confidence of that blob to the same remains low. Thus the four Boolean predicates for these occlusion primitives can be constructed as follows.

$$\epsilon(I)[j,t] = \exists i[\Theta_{AF}[j,i](t) = 1] \wedge [\Theta_F[i](t) = 1] \quad (3)$$

$$\epsilon(D)[j,t] = [\Theta_A[j](t) = 0] \wedge [\Psi_A[j](t) = 0] \quad (4)$$

$$\epsilon(P)[j,t] = \forall i[\Psi_{FA}[i,j](t) = 1] \wedge [\Theta_F[i](t) = 1] \wedge [\Psi_A[j](t) \geq 1] \quad (5)$$

$$\epsilon(C)[j,t] = \exists i[\Theta_{AF}[j,i] = 1] \wedge [\Theta_F[i](t) > 1] \quad (6)$$

Updates are applied to color, shape and trajectory of individual agents under $\epsilon(I)$, and only to , the trajectory of agents under $\epsilon(P)$ and $\epsilon(C)$. Agents under $\epsilon(D)$ are moved from the active set to the putative set. This enables the system to remain updated with agent features while keeping track of them.

The **entry/reappearance** of an agent is attributed to the existence of a foreground blob $F_i(t)$ in the scene having no association with any agent from $\mathcal{S}_A(t-1)$ and the corresponding Boolean predicate is constructed as $\text{NEWBLOB}_i(t) = [\Theta_F[i](t) = 0] \wedge [\Psi_F[i](t) = 0]$. The features of the new blob $F_i(t)$ are matched against those in $\mathcal{S}_P(t-1)$ to search for the reappearance of agents. If a match is found, the agent is moved from $\mathcal{S}_P(t-1)$ to $\mathcal{S}_A(t)$. Otherwise, a new agent is added to $\mathcal{S}_A(t)$. Similarly, an agent is declared to **exit** the scene, if its motion predicted region lies outside the image region and is thus removed from the active set. The (re)appearance of a **new**

agent (blob) in an attentional window $\omega_j(t)$ around the j^{th} agent is represented as the Boolean predicate $\epsilon(N)[j,t] = \exists i[\text{NEWBLOB}_i(t)] \wedge [\gamma(F_i(t), \omega_j(t)) \geq \eta_N]$. The region $\omega_j(t)$ is typically constructed as a $2s_x \times 2s_y$ rectangle where s_x and s_y are the respective width and height of the minimum bounding rectangle of $A_j(t)$. The threshold η_N is empirically set to 0.75 to ensure significant fractional overlap with the attentional window. It is worth noting, that $\epsilon(N)[j,t]$ can occur in conjunction with one of $\epsilon(I)/\epsilon(P)/\epsilon(C)$ and is expressed accordingly.

3 Activity Discovery

Activities involving multiple agents are identified based on temporal sequences of occlusion primitives $\mathcal{E} = \{\epsilon_r\}_{r=1}^{R}$ that constitute an *activity* involving the j agents. Temporal sequence mining involves finding paths in an *activity tree* $\mathcal{T}_\alpha$. An empty (first in first out) buffer β_j of length L and the activity tree $\mathcal{T}_\alpha(j)$ is initialized with a root node ρ_j at the very first appearance of every j^{th} agent in $\mathcal{S}_\mathcal{A} \cup \mathcal{S}_\mathcal{P}$. This ensures the discovery of variable length event sequences whose length do not exceed L. Each node of $\mathcal{T}_\alpha(j)$ is a two tupple $\mathcal{T}_n \equiv (\epsilon, \pi)$ containing the primitive $\epsilon \in \mathcal{E}$ and a real number $\pi \in (0, 1]$ signifying the probability of occurence of the path $\{\rho_j, \ldots, \mathcal{T}_n\}$.

Let, $\beta_j[l](t)$ be the o-primitive at the l^{th} depth of the buffer, the most recent one being logged at $l = 0$ and the last one at $l = L - 1$. The event primitive $\epsilon(j,t) \in \mathcal{E}$ detected for the j^{th} agent at the t^{th} instant is pushed to β_j, iff the current event primitve changes from that of the previous, i.e. $\epsilon(j,t) \neq \epsilon(j,t-1)$. This prevents learning the variable length temporal sequences of similar events as separate activites. Let, $B^{(l)}(j,t) = \{\alpha_u^{(l)}(j,t)\}_{u=1}^{b_l}$ be the set of l-length paths (originating from ρ_j) of $\mathcal{T}_\alpha(j,t)$. 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,

$$\pi_u^{(l)}(j,t) = (1 - \eta_l(t))\pi_u^{(l)}(j,t-1) + \eta_l(t)\delta(u-b) \quad (7)$$

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 $\eta_l(t) = max\left(\frac{1}{t}, \eta\right) \forall l$. A new event primitve is added to the tree with an intial probability of $\eta_l(t)$ and the self normalizing nature of equation 7 ensures the properties of the probability measure at each depth of $\mathcal{T}_\alpha(j)$. It is worth noting that this procedure is closely related to offline sequence mining [1] and VLMM learning [9].

Semantic labels can be assigned to different sequences in the occlusion-primitive space, and subsequences may constitute sub-activities. For example, for the **h**iding activity (Figure 3), the activity tree is learned with $L = 10$ and $\eta = 0.01$ and the sequences $(\epsilon(I))$, $(\epsilon(P))$, $(\epsilon(D))$, $(\epsilon(I), \epsilon(P))$, $(\epsilon(P), \epsilon(D))$, $(\epsilon(D), \epsilon(P))$, $(\epsilon(P), \epsilon(I))$, $(\epsilon(I), \epsilon(P), \epsilon(D))$, $(\epsilon(P), \epsilon(D), \epsilon(P))$, $(\epsilon(D), \epsilon(P), \epsilon(I))$, $(\epsilon(I), \epsilon(P), \epsilon(D), \epsilon(P))$, $(\epsilon(P), \epsilon(D), \epsilon(P), \epsilon(I))$ and $(\epsilon(I), \epsilon(P), \epsilon(D), \epsilon(P), \epsilon(I))$ are obtained as the paths in the learned tree. Sub-sequences that have been learned as part of this process may be labelled **going to hide**, **coming out of hiding** and **hiding and reappearing** as the variable length sequences $(\epsilon(I), \epsilon(P), \epsilon(D))$, $(\epsilon(D), \epsilon(P), \epsilon(I))$ and $(\epsilon(I), \epsilon(P), \epsilon(D), \epsilon(P), \epsilon(I))$ respectively.

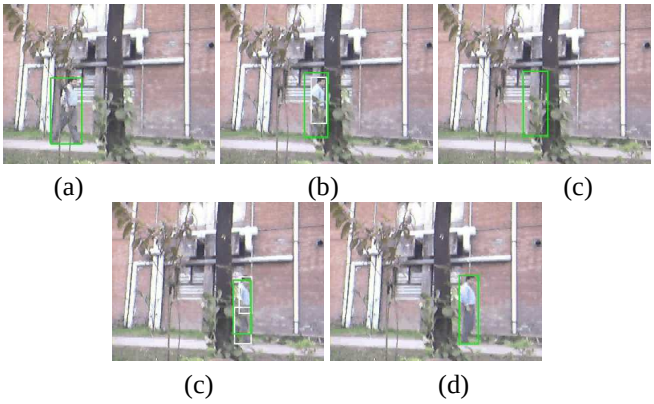


Figure 3. Discovering the hiding activity. (a) Agent approaching tree; (b) Agent partially occluded by tree; (c) Agent hiding behind tree; (d) Agent partially occluded while reappearing; (e) Agent reappears completely

From this example, we observe the efficiency of incremental sequence learning for the discovery of activities of variable temporal durations. More so, the event primitives themselves are leanred as atomic activities as unity length sequences. However, it is worth noting that all learned sequences do not necessarily corrrespond to meaningful activities to a human observer in a particular surveillance scenario. Pruning the tree with an information theoretic measure [4] or a probability threshold does not solve this problem as the discovered sequences with high frequency may not signify a meaningful activity in a specific application domain, although a rare event might be important. Thus a database of meaningful sequences should be constructed by user interaction for future application specific activity recognition purposes. Further results of activity discovery from different video sequences are presented in Section 4.

4 Testing on Complex Data

The proposed algorithm has been tested on a varitey of outdoor surveillance videos consisting of single and multi-

person activity, involving open spaces, woods, and traffic situations. The results of o-primitive assisted multi-person tracking are shown in figures 3 and 4. We choose $\eta_A = 0.6$, $\eta_F = 0.6$ achieving an overall tracking accuracy of 82.54%. For all the experiments, the activity trees are learned with $L = 5$ and $\eta = 0.01$. The proposed methodology for activity discovery from the results of o-primitive assisted multi-agent tracking is performed at an approximate rate of 6.0 frames per second while operating on 320×240 color images on a $1.6GHz$ Pentium-4 PC.

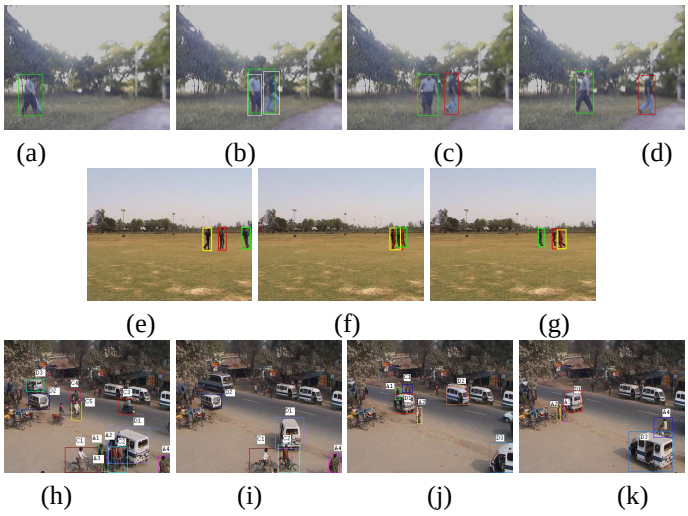


Figure 4. (a)-(d) Simple Splitting: transitions between isolated, fragmented, new blob emergence. (e)-(g) Joining a group and separating; (h)-(k) Complex multi-agent interactions while embarking (h,i) and disembarking (j,k) from a vehicle - isolated vehicle followed by multiple vehicle-person mergers/splits.

The activity of **splitting** is discovered from the sequence shown in figures 4(a)-(d), where the agents enter together and separate in the field of view. The mean-shift tracker initially converges on both of them (figure 4(b)) as they tend to separate, which is detected as a partial occlusion for a few frames depending on the relative velocity between the agents. Later, the tracker converges on one of them and a new agent is formed in its neighborhood. The activity tree is found to learn the sequences $(\epsilon(I), \epsilon(P))$, $(\epsilon(P), \epsilon(I) \wedge \epsilon(N))$, $(\epsilon(I) \wedge \epsilon(N), \epsilon(I))$, $(\epsilon(I), \epsilon(P), \epsilon(I) \wedge \epsilon(N))$, $(\epsilon(P), \epsilon(I) \wedge \epsilon(N), \epsilon(I))$ and $(\epsilon(I), \epsilon(P), \epsilon(I) \wedge \epsilon(N), \epsilon(I))$ apart from the unity length ones. The longest sequence $(\epsilon(I), \epsilon(P), \epsilon(I) \wedge \epsilon(N), \epsilon(I))$ is found to tally with the process of splitting. The subsequences signifying the parts of the splitting process are rejected in the process of user interaction.

Similarly, the activities of **forming a group**, **separating from a group** and **forming a group and separating** (figure 4(e)-(g)) are learned as $(\epsilon(I), \epsilon(C))$, $(\epsilon(C), \epsilon(I))$ and $(\epsilon(I), \epsilon(C), \epsilon(I))$ respectively, the former ones naturally being subsequences of the entire process.

The scene in figure 4(h)-(k) shows a complex interaction between several agents, some of whom are arriving on bicycles and boarding a vehicle (this type of vehicle is known as "tempo"). While embarkation and disembarkation scenarios are easily identified, much more can be done. For each tracked agent a shape and motion history is available, which can potentially be used to obtain clusterings in the agent space. This possibility is highlihted in the labels used in these figures which distinguish several categories (e.g. A = person, C = man+bicycle, D=tempo). This would enable us to classify a crowd to be either homogeneous (e.g. all human beings) or heterogeneous (e.g. human being(s) and vehicle(s)) and the respective event primitives for an agent in the crowd can then be denoted as $\epsilon(C_m)$ and $\epsilon(C_r)$. The sequences $(\epsilon(I), \epsilon(C_m), \epsilon(C_r), \epsilon(D))$ and $(\epsilon(I), \epsilon(C_r), \epsilon(D))$ are found to correspond to the cases of **embarking a vehicle in a group** and **embarking a vehicle alone** respectively. **Disembarking** on the other hand, is simply discovered as a process of splitting.

5 Conclusion

This paper demonstrates the power of a set of occlusion features called o-primitives, and temporal sequences of o-primitives are posited as a powerful tool for identifying multi-agent interactions. We outline the temporal mining process for detecting event hierarchies in this space. A variety of composite events can be distinguished based merely on the occlusion phenomena.

The o-primitives constitute only one of the many tools available for activity recognition. Even without extensive camera calibration or domain knowledge, a further characterization of the qualitative motion in the image-space may include *translate left/right/towards/away*, *rotate*, *speeding up*, *halting* etc. which can by themselves be informative in certain domains.

Complex activities are characterized by the type-of-activity (predicate) as well as the ordered set of agents participating in it, as well as other characteristics such as time, place, manner etc. (arguments). In this domain-independent presentation we have not focused on agent shape or ground-plane based motion characterization, but as shown in figures 4(h)-(k), clustering the agents into classes based on temporal characterization of motion and shape behaviours would immediately provide a set of arguments for the predicate, the ordering of which can also be inferred based on the temporal sequence. Other quantitative measures associated with the event such as time, location, manner, etc can also be inferred from the visual sequence. One of the future goals of this work is to unify these motion primitives,

spatial tags and quantifiers to mine a richer set of activites leading to richer recognition semantics.

References

[1] R. Agrawal, T. Imieliand, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216. ACM Press, 1993.

[2] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition*, volume 2, pages 142–149, 2000.

[3] A. Galata, A. G. Cohn, D. Magee, and D. Hogg. Modeling interaction using learnt qualitative spatio-temporal relations and variable length markov models. In *European Conference on AI (ECAI)*, July 2002.

[4] A. Galata, N. Johnson, and D. Hogg. Learning variable-length markov models of behavior. *Computer Vision and Image Understanding*, 81(3):398–413, 2001.

[5] D. Gavrilla. Visual analysis of human movement: A survey. In *Computer Vision and Image understanding*, volume 73, pages 82–98, 1999.

[6] G. Granlund. Organization of architectures for cognitive vision systems. In *Proceedings of Workshop on Cognitive Vision*, Schloss Dagstuhl, Germany, October 2003.

[7] P. Guha, D. Palai, K. Venkatesh, and A. Mukerjee. A multiscale co-linearity statistic based approach to robust background modeling. In *Seven'th International Asian Conference on Computer Vision*, January 2006.

[8] I. Haritaoglu, D. Harwood, and L. Davis. W4 : Real time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:809–830, August 2000.

[9] D. Ron, Y. Singer, and N. Tishby. The power of amnesia: learning probabilistic automata with variable memory length. *Machine Learning*, 25(2-3):117–149, 1996.

[10] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environments. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 406–413, July 2004.