Hybrid Hierarchical Learning from Dynamic Scenes

Prithwijit Guha[1], Pradeep Vaghela[1], Pabitra Mitra[2], K.S. Venkatesh[1], and Amitabha Mukerjee[2]

[1] Department of Electrical Engineering, Indian Institute of Technology, Kanpur, Kanpur - 208016, UP, India
{pguha, pradeepv, venkats}@iitk.ac.in
[2] Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur, Kanpur - 208016, UP, India
{pmitra, amit}@cse.iitk.ac.in

Abstract. The work proposes a hierarchical architecture for learning from dynamic scenes at various levels of knowledge abstraction. The raw visual information is processed at different stages to generate hybrid symbolic/sub-symbolic descriptions of the scene, agents and events. The background is incrementally learned at the lowest layer, which is used further in the mid-level for multi-agent tracking with symbolic reasoning. The agent/event discovery is performed at the next higher layer by processing the agent features, status history and trajectory. Unlike existing vision systems, the proposed algorithm does not assume any prior information and aims at learning the scene/agent/event models from the acquired images. This makes it a versatile vision system capable of performing in a wide variety of environments.

1 Introduction

In recent years, there has been an increasing interest in developing cognitive vision systems capable of interpreting the high level semantics of dynamic scenes. A good overview of cognitive vision system architectures can be found in [1]. Traditional approaches to dynamic scene analysis operate only in restricted environments with predefined and/or learned quantitative object and behavior models. Such models are often fragile and lead to modeling errors; thereby degrading the performance in most practical situations. A hybrid multi-layered vision architecture, consisting of both quantitative and qualitative models which are more robust and immune to modelling errors, essentially processes the visual data at lower levels and extracts worthwhile semantic information for analysis in higher layers.

This work proposes a multi-layered cognitive vision system for agent/event discovery from dynamic scenes, which processes information at various stages of knowledge abstraction, each layer deriving its percept model from the observations obtained from lower level(s). The system initializes with a few preset capabilities involving scene feature (color and shape) extraction, and applies multi-agent

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 212–217, 2005.

tracking with symbolic reasoning, unsupervised agent categorization and event discovery with variable length Markov models. Raw visual data is processed at the lowest level to yield a perception of the background model of the scene. A hybrid analysis involving symbolic reasoning and feature extraction from the image data is performed at the mid-level for multi-agent tracking. The higher layer essentially categorizes the quantitative agent features and qualitative event descriptors in an unsupervised learning framework. The system can be further extended into upper layers depending on the application context, which typically requires user interaction for tagging learned categories and generating linguistic descriptions. Figure 1 shows the functional architecture of the proposed framework.

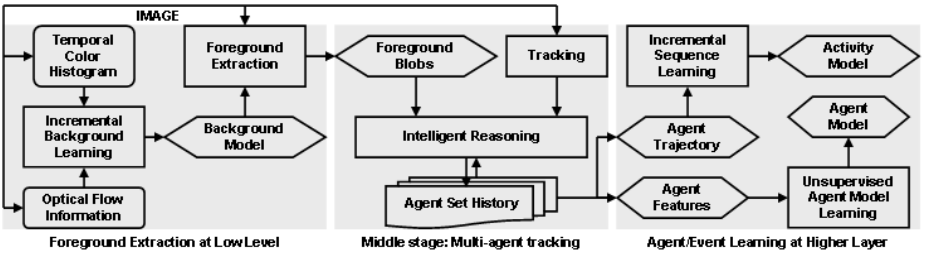

Fig. 1. The Proposed hierarchical scene analysis architecture

The paper is organized in the following manner. Section 2 explains the processing at the lower level of the proposed system. The symbolic reasoning for multi-agent tracking at the mid-level are explored in Section 3. Agent categorization and event discovery are described in Section 4. The results of experimentation are briefly described in Section 5. Finally, we conclude the paper in Section 6.

2 Background Learning for Low Level Vision

Traditional vision systems with an object centered approach learn the background with a specific signal model from a few initially unintruded (pure background) frames. However, in most practical cases, signal models do change and pure background frames can't be availed of for training. Thus, a view centered approach is to be adopted for estimating the background model in an adaptive learning framework. The usual approach to incremental background modeling involves fitting (temporally evolving) Gaussian mixture models [2] on the temporal pixel color histogram. Recently, Gutches et al. [3] have proposed an online background learning algorithm, which combines the temporal histogram features along with optical flow information leading to improved modeling performance. This approach is adopted in our work for modeling the background $\mathbf{B}_t$ learned till the t^{th} instant. This is used to perform the background-foreground segmentation of the image Ω_t followed by a connected component analysis to generate the set $\mathcal{F}_t = \{F_i(t)\}_{i=1}^{n_t}$ of disjoint foreground blobs. The extracted foreground blobs are used for symbolic reasoning to track multiple agents at the mid-level of visual

(a) (b) (c)

Fig. 2. Background learning. (a) Sample traffic scene, (b) Background learned after 100 frames, (c) Foreground blobs extracted from (a) using (b).

processing. The results of foreground extraction with incremental background learning (from a traffic video sequence) are illustrated in figure 2.

3 Symbolic Reasoning for Multi-agent Tracking

Agent/event discovery primarily depends on the availability of reliable features and is often challenged by occlusions arising out of crowding and obstructions by background objects. Unlike conventional object oriented approaches [4] to occlusion handling that assume prior shape and motion models and also a ground plane, we propose a reasoning scheme that is not restricted by specific agent/environment models and detects several event primitives. This assists the learning process in selective agent feature updates. [5] provides a detailed discussion of multi-agent tracking which we summarize here.

Intelligent reasoning 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 $\mathcal{S}_P(t)$ of agents of which the system has lost track. The system initializes itself with empty sets and the agents are added (removed) as they appear (disappear) in (from) the field of view. The j^{th} agent in the active set 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 [6] initialized with the motion predicted position from the trajectory $\{C_j(t-t')\}_{t'=1}^{\tau}$ are used to localize $A_j(t)$ in the t^{th} frame. To associate the j^{th} agent with the i^{th} foreground blob, we construct the thresholded *localization confidence matrix* $\Theta_{AF}(t)$ and the *attribution confidence matrix* $\Psi_{FA}(t)$. Measures of foreground regions per agent ($\Theta_A[j](t)$ and $\Psi_A[j](t)$) and agents per foreground region ($\Theta_F[i](t)$ and $\Psi_F[i](t)$) can be computed from these matrices.

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

$$\Theta_A[j](t) = \sum_{i=1}^{n_t} \Theta_{AF}[j,i](t), \qquad \Theta_F[i](t) = \sum_{j=1}^{m_{t-1}} \Theta_{AF}[j,i](t) \quad (2)$$

$$\Psi_A[j](t) = \sum_{i=1}^{n_t} \Psi_{FA}[i,j](t), \qquad \Psi_F[i](t) = \sum_{j=1}^{m_{t-1}} \Psi_{FA}[i,j](t) \quad (3)$$

We construct four Boolean predicates: $\textsc{IsoUnOccTrk}_j(t)$ for agent *isolated, unoccluded and well tracked*; $\textsc{LostTrack}_j(t)$ for agent *lost track* of; $\textsc{IsoPartOcc}_j(t)$ for agent *partially occluded*; $\textsc{Crowd}_j(t)$ for agent in a *crowd*.

$$\textsc{IsoUnOccTrk}_j(t) = \exists i[\Theta_{AF}[j,i](t) = 1] \wedge [\Theta_F[i](t) = 1] \wedge [\Psi_F[i](t) = 1] \quad (4)$$

$$\textsc{LostTrack}_j(t) = [\Theta_A[j](t) = 0] \wedge [\Psi_A[j](t) = 0] \quad (5)$$

$$\textsc{IsoPartOcc}_j(t) = \forall i[\Psi_{FA}[i,j](t) = 1] \wedge [\Psi_F[i](t) = 1] \wedge [\Psi_A[j](t) > 1] \quad (6)$$

$$\textsc{Crowd}_j(t) = \exists i[\Theta_{AF}[j,i] = 1] \wedge [\Theta_F[i](t) > 1] \quad (7)$$

(a) (b) (c) (d)

Fig. 3. Cases of occlusions. (a-b) Partial occlusions: agent detected as multiple foreground blobs; (c-d) Crowding: multiple agents merge to form a single blob.

Color, shape and trajectory of individual agents under IsoUnOccTrk, but only the trajectory of agents under IsoPartOcc and Crowd are continuously updated. Agents under LostTrack are moved from the active set to the putative set.

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 $\mathcal{S}_A(t-1)$. Hence, the corresponding Boolean predicate $\textsc{NewBlob}_i(t)$ is,

$$\textsc{NewBlob}_i(t) = [\Theta_F[i](t) = 0] \wedge [\Psi_F[i](t) = 0] \quad (8)$$

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's entry is declared and 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 vision system often encounters the phenomenon of **splitting**, when two (or more) agents enter the scene in a group and separate later within the field of view. In such cases, they are initially learned as a single agent and the split is eventually detected as a fragmentation (such as that caused by partial occlusion) for the first few frames (the exact number depends on the relative velocity between separating agents). Afterwards, the tracker converges on one agent and loses track of the other(s), which eventually emerge as a new region(s) and is (are) added as new agent(s).

4 Discovery of Agents and Events

The agents are essentially represented by their color, shape and motion features. The instances of the same class can have significantly different color and motion

features, thereby leaving the shape as a more reliable descriptor. In this work, we opt for the second to seventh order Zernike moments [7], which serve as a shape feature. The shape feature vector $X_j(t)$ of the j^{th} agent is only computed when the agent is isolated and well localized and a Gaussian mixture model is incrementally learned over this feature set. Empirically, the selected feature set seems to provide consistent classification which agrees well with that of a human observer. The categorization algorithm has successfully classified the instances of *man*, *man on bike*, *vehicle* and *rickshaw* (figure 4(a)-(d)).

Events are of two different categories, viz. *actions* (characteristic state space trajectories of individual agents) and *interactions* (activities involving multiple agents). Typical examples of actions are the path taken by a vehicle in a traffic scenario or the pose sequence exhibited by a dancer. The events of chasing, overtaking of vehicles refer to interactions in a homogeneous group and the examples for a heterogeneous group include activities like boarding a vehicle, riding a bike etc. Variable length Markov models (VLMMs) have been used previously for the purpose of learning actions [8] and interactions [9]. In this work, we apply the VLMMs for online event discovery by learning the frequent sequences of symbolic descriptors acquired earlier. A detailed discussion on learning with VLMMs can be found in [8].

Fig. 4. Instances of learned agents. (a) Man, (b) Man on Bike, (c) Vehicle, (d) Rickshaw. Discovered multi-agent interactions. (e-f) Person on cycle detected followed by person-cycle split: person getting off a cycle, (g-h) Person and cycle isolated followed by person-cycle crowded: The event of person embarking a bicycle, (i-j) A heterogeneous crowd followed by track loss of several agents: people embarking a vehicle, (k-l) Isolated vehicle followed by vehicle-person split: The event of disembarking from a vehicle.

5 Results

The proposed algorithm for event discovery from dynamic scenes is applied to a traffic scenario. The agents and events are discovered online from a video sequence acquired with a static camera. The multi-agent tracking algorithm provides us with symbolic descriptors of object states at each time instant. We assume the interacting agents to be present within a specific attentional window. spatio-temporal interactions are learned as frequently co-occurring sets of agent states. Results show the discovery of events like a *man boarding a bicycle*, *man disembarking the bicycle*, *people boarding vehicle* and *people getting off from vehicle*. The results of discovered sequences are shown in figure 4.

6 Conclusion

We have proposed a hierarchical architecture for discovering events from dynamic scenes. Symbolic and sub-symbolic learning is performed at different processing levels in the hierarchy. The architecture requires minimal prior knowledge of the environment and attempts to learn the semantic informations of dynamic scenes. Qualitative reasoning at a higher level induces tolerance to learning error at lower levels . This makes it suitable for a wide range of applications.

References

1. Granlund, G.: Organization of architectures for cognitive vision systems. In: Proceedings of Workshop on Cognitive Vision, Schloss Dagstuhl, Germany (2003)
2. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Volume 2., IEEE Computer Society (1999) 246–252
3. Gutchess, D., Trajkovics, M., Cohen-Solal, E., Lyons, D., Jain, A.K.: A background model initialization algorithm for video surveillance. In: Proceedings of Eighth IEEE International Conference on Computer Vision. Volume 1. (2001) 733–740
4. Haritaoglu, I., Harwood, D., Davis, L.: W4 : Real time surveillance of people and their activities. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 809–830
5. Guha, P., Mukerjee, A., Venkatesh, K.: Efficient occlusion handling for multiple agent tracking with surveillance event primitives. In: The Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. (2005)
6. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Transaction on Pattern Analysis Machine Intelligence **25** (2003) 564–575
7. Teague, M.R.: Image analysis via the general theory of moments. Optical Society of America **70** (1980) 920–930
8. Galata, A., Johnson, N., Hogg, D.: Learning variable-length markov models of behavior. Computer Vision and Image Understanding **81** (2001) 398–413
9. 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