A SYNERGISTIC MODEL FOR INTERPRETING HUMAN ACTIVITIES AND EVENTS FROM VIDEO : A CASE STUDY

N. Bourbakis [1,2,3], G. Bebis [4], J.Gattiker [5]

1- AIIS Inc, Vestal NY, 2-Binghamton Univ., Dept.EE, IVMV Lab, Binghamton NY 13902, 3-Technical University of Crete, ECE Dept. Chania 73100, Crete, Greece, 4-Univ. of Nevada-Reno, CS Dept. Reno, NV 89557, 5-Los Alamos-national Labs, Los Alamos, NM 87544

Abstract

This paper describes a new approach for representing, recognizing and interpreting human activity from video. The approach presented (at the conceptual level) here is a model based on the hierarchical synergy of three other models (the L-G graph, the SPN graph and a NN model). In particular, in our project human activity is strongly related with the ability of describing and interrelating events. Thus, the L-G graph provides a powerful description of the structural image features presented in an event, the SPN model offers a description of the functional behavior of the changes or operations in video presented in an event, and the NN model provides the capability of extracting and learning behavioral patterns, presented in human activities.

1. INTRODUCTION

In the recent years the scientific accomplishments in image understanding and visual languages fields have been shifted into the video understanding domain. Several research efforts have been presented in the literature to achieve objects motion in a sequence of images, and some cases basic understanding of human behavior, such as description of a human that carries a box, a human that walks, etc.[1-18]. In particular, relative research work has been done on trajectory guided tracking and recognition of actions [14], stochastic temporal models of human activities [13], bayesian approach to human activity recognition [12], complex visual activity recognition using a temporally ordered DB [11], layered probabilistic recognition of human action [10], learning and recognizing human dynamic in video sequences [9], probabilistic recognition of human actions [8], recognizing human action in time-sequential images using HMM [7], visual understanding of dynamic hand gestures [15] etc. These methodologies, however, provide neither a complete interpretation of an event nor the cause of the event. Thus, this paper targets the area of human activity interpretation and recognition with the purpose of developing methods that would cue human experts to activities that may be of interest, while ignoring the vast amount of irrelevant information. Specifically, our emphasis is on amplifying the capabilities of humans in the surveillance system. This assistance can take different forms, such as cueing attention to unusual or proscribed events to allow a human observer to efficiently monitor more information; (such tireless observation can lower the possibility that an event of potential interest may go unnoticed. Robust detection and tracking of humans is a challenging problem due to the changeable morphology of the human body. Complex human movements (e.g. walking ,running) can give rise to widely different human appearances. The problem becomes even more difficult when considering multiple moving objects, which might occlude or destruct the motion of each other, and complex non-static natural backgrounds. This paper addresses the need to track humans and objects in a visual scene, and assess the activities in the scene. Starting from low-level vision operations of segmentation and contour generation, characterization of the scene objects as graph descriptions will allow recognition and characterization of the objects. From object characterization, a correspondence from structural object features to functional attributes will result in a description of the activity. The model proposed here is based on the implementation of three main components: (1) segmentation and classification of moving objects, (2) tracking of humans, and (3) recognition of human activity.

2. THE SYNERGISTIC MODEL

It is known that an event is the result of a sequence of actions, and it holds structural and functional information. The recognition of a human activity is strongly related with the ability of describing and interrelating events. Thus the capability of describing, analyzing, recognizing events leads to events understanding and may analyzing and predicting human activities. Thus, the idea behind this paper here is a synergistic methodology based on Local-Global (L-G) graphs, Stochastic Petri-nets (SPN) and Neural nets (NN). The L-G graph provides a powerful description of the image structural features presented in an event, the SPN model offers a description of the functional behavior of the changes or operations in video presented in an event, and the NN model provides the capability of extracting and learning behavioral patterns, presented in human activities. These models work in a synergistic hierarchical manner, where the advantages of one model at a particular level are employed by the model of the next level. More specifically, the L-G graph maps its structural descriptions into the SPN model by converting it into a Stochastic Petri-net graph (SPNG) that holds both structural and functional representations. Due to high complexity (a great volume of states for each object) at the SPNG level, a search process becomes difficult and time consuming. Thus, by mapping the SPNG features into a NN model the complexity problem becomes tractable based on the ability of the neural nets to deal with a great volume of data. This mapping is possible via an isomorphism of the SPNG onto a Neural Net for the extraction of the events interpretation and learning from them. *Note that some important features of this synergistic model are the ability of the neural part to learn certain sequences of actions and predict the output before their completion, also the ability to analyze a sequence of actions and determine the possible "causes".*

2.1. MULTIPLE CAMERAS

To implement a system that understands human activity, the viewing system should be able in general to image the tracked human(s) in a broad area over a long period of time. Using a fixed single camera has limitations since it restricts tracking to a very narrow area due to the restricted viewing angle of the system. A moving camera with some degree of rotational freedom increases the viewing angle, however, it complicates the implementation by adding the motion estimation of both the viewing system and the subject of interest. In the system described here, multiple fixed cameras mounted in the area of interest to track and monitor the motion of individuals in sequences of monocular images are used. Since occlusion is view angle specific, multiple cameras will reduce the chance the occlusion is present in all views. Multiple cameras will also alleviate the difficulty when certain views are confused. To improve the robustness of the segmentation and tracking, the system will also use information from multiple calibrated cameras. Although the fusion of information from multiple cameras will improve the segmentation and tracking, there will still be unresolved cases. To further improve the results, the system to be implemented will not be strictly feed-forward, from low-level operators to high level recognition processes. On the contrary, feedback from the recognition process will also guide the segmentation and tracking processes.

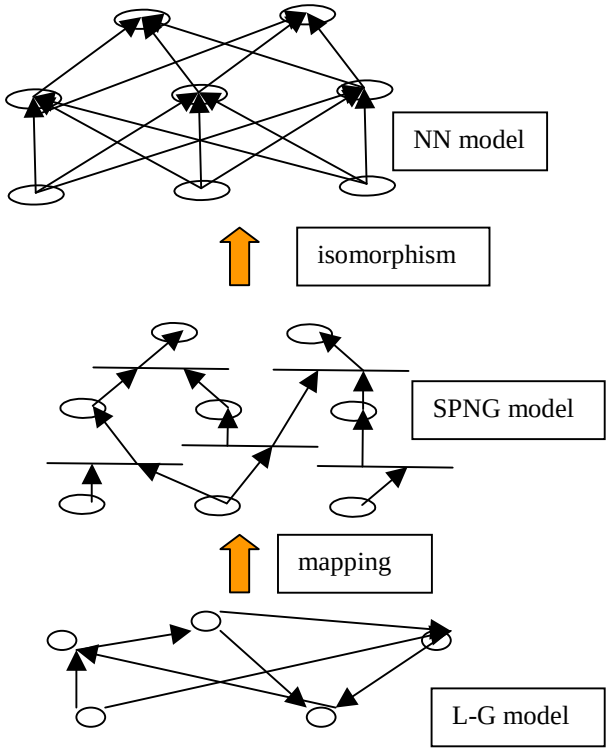


Fig. 1: The synergistic model

2.2. THE L-G GRAPH

The L-G graph is capable of describing with accuracy and robustness the features contained in an image. It consists of two parts. The local (LG) region graph and the global (GG) image graph. A brief description is provide bellow.

THE REGION OR LOCAL GRAPH

A Local graph G holds information of a contour –line of an image region after segmentation.

$\mathbf{G = N_1 a^c_{12} N_2 a^c_{23} N_3 \ldots N_k a^c_{k1} N_1 \otimes N_i\ a^p_{ij} N_j \otimes \ldots \otimes N_n a^{rd}_{nm} N_m}$ ...

where, ⊗ represents the graph relationship operator, and each **Ni** maintains the structural features of the corresponding line segment, thus, **Ni** = { sp, orientation (o), length (le), curvature (cu)}, and **aij** holds the relationships among these line segments, thus, $\mathbf{a_{ij}}$ = { connectivity (c) , parallelism (p) , symmetry (s), relative magnitude (rm), relative distance (rd), etc}

The missing elements for a global visual perception of an image are: the color (or texture) of each region, its relative geographic location (distance and angle) among the other regions, its relative size in regards with the other regions, etc. One way to obtain these additional features is the development of the global image graph GG.

THE IMAGE GLOBAL GRAPH

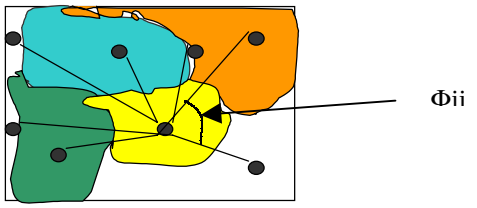
The global image graph attempts to emulate a human-like understanding by developing global topological relationships among regions and objects. More specifically, for each image region Mi, a skeletonization task is performed and the final center of gravity GCg(i,x,y) is defined [N3].

When all the final centers of gravity have been defined for every image region, the global image graph is developed

$$\mathbf{GG(Ak) = (P_1 R_{12} P_2)\ \Phi_{23}\ (P_1 R_{13} P_3) \ldots (P_1 R_{1n-1} P_{n-1})\ \Phi_{n-1n}\ (P_1 R_{1n} P_n)}$$

where **Pi** is a node that represents an image region graph, its color, and its GCg(i,x,y), **Rij** represents the relative distance between two consecutive Gcg, and the orientation of each dg, **Φij** represents the relative angle between consecutive distances dg(i) and dg(j).

Image regions and the GG graph



$GGA_{(N1)}) = (P_1 R_{12} P_2)\ \Phi_{23}\ (P_1 R_{13} P_3)\ \Phi_{34}\ (P_1 R_{14} P_4)$..

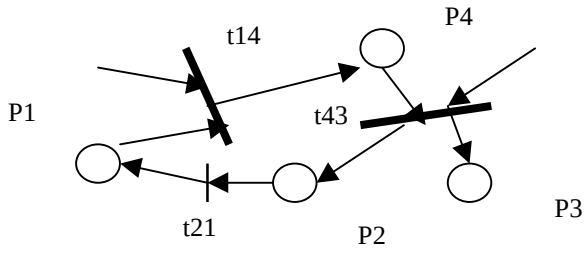
An important feature of the L-G graph is its ability to describe 3-D scenes. The only difference between 2-D from 3-D is that in 3-D the local graph will represent 3-D surfaces and the global graph will appropriately interrelate them .

2.3. THE SPNG MODEL

The graph models mentioned above have the capability of holding structural information about targets. Thus, the missing element is the functional behavior of a target. The functional behavior of a target is described by the states in which a particular target could be transferred after an appropriate triggering. A successful and power model capable of describing (or modeling) the functional behavior of a system is the Stochastic Petri-net (SPN) model. Thus, in order to maintain the structural features of the graph model and the functional features of the SPN model, a mapping is presented here, where the SPN model is transformed into a SPN graph model as follows [GN]:

m : G → SPNG

where, **{Ni} → {Pi},** graph-nodes correspond into SPN places , and **{aij} → {tij},** relationships corresponds into SPN transitions.



An SPNG model

The SPNG above illustrates a target that has four different states (Places Pi, i=1,2,3,4). Each place Pi has its own structural features transferred from the corresponding graph node Ni. The transitions t14 and t43 represent relationships among the same parts of a target and a stochastic distribution of time required to fire that transition. The t21 requires no time to fire.

2.4. THE ISOMORPHISM OF SPNG ONTO NN

The SPNG model has to deal with a great volume of information due to many functional states associated with each object. Thus, the interpolation and correlation of such amount of information becomes difficult and time consuming. An alternative solution is the use of a neural network model. The neural model alone , however, is not able to efficiently deal with structural representation of knowledge. A solution proposed here is the transferring of the SPNG structural and functional features into a NN via an isomorphism: **I : SPNG → NN,** where the places Pi → ni (ni is a neuron-nodes of a NN), and transitions tij → wi, (wi represents a threshold logic net).

RECOGNITION AND LEARNING

We have previously described how specific procedural activities can be incorporated into our system for activity recognition in scene situation analysis, as well as the assessment of the results of activity in constrained situations. We will also investigate and demonstrate the capability of this framework to learn normal and abnormal patterns from the data. Given a time-sequence of humans, their tracks, and a description of their position and activities, we propose to use neural adaptive system to characterize and learn normal behaviors, and thus also recognize anomalous or unusual behaviors. Starting from a tracking record of humans and their activities (object manipulation, etc.), normal profiles can be built up at various levels of description complexity, from simple path tracking based on Markov Models, to complex characterizations of sequential activities using time-series and spatial descriptions of arbitrary complexity, such as in neural models. These models can then be used to isolate unusual, suspicious, or alarming behavior from the identified scenes. The isomorphism between these descriptions learned dynamically in a neural formalism, and the structural and functional descriptions isolated through the scene recognition provides the crucial link for combining learning capabilities of the adaptive systems with the human and object description structures that will be developed for underlying scene feature recognition.

3. METHODS AND TASKS

3.1. DETECTION OF MOVING OBJECTS

The system will be initialized by acquiring measurements of the scene over a number of video frames. The goal is to build a powerful representation of the background which will allow us to extract the moving objects in the scene more robustly. Towards this objective, we propose using an eigenspace representation of the background. What is interesting about this representation is that when an image containing new objects is projected onto the background eigenspace, then its reconstructed counterpart does not contain the new objects anymore. This representation needs to be updated over time to account for changes in lighting conditions and new objects in the scene. In a long term, this eigenspace will describe the range background appearances that have been observed. The key to using the eigenspace background representation is on our ability to update it efficiently over time. We will deploy recent results in numerical mathematics (recursive Singular Decomposition techniques) to implement incremental approaches that will allow us to update the eigenspace of the background in real time. To further improve the segmentation results, we will use (i) fusion from segmentations obtained using other viewpoints (multiple cameras), (ii) information from the tracking component (predicted locations), and (iii) feedback from the high-level human activity recognition component.

3.2.TARGET DETECTION, EXTRACTION AND REPRESENTATION

Connected components analysis, and morphological operations will be used to improve segmentation quality. Each segmented object will then be divided into a number of

regions which will be used for building its L-G graph. Specifically, each human will be divided into a number of regions using color, texture, and motion information. Each region will be represented by a mixture of Gaussian learned by using the EM algorithm. In this task the automated detection, extraction and representation in graph forms of a variety of targets, such as human, objects, animals will be performed. In particular, each segmented image frame will be described by the Local-Global (L-G) graph model. The L-G graph model provides a flexible representation of the targets and their surrounding in an image frame independent from rotation and shifting. The target detection problem presents two possible scenarios: a) the targets are known to the system, and b) the targets are unknown. For the first scenario , the target is known to the system, it means that its L-G model is available. Thus, the system searches for the target L-G graph form at the image L-G graph representation by using a fuzzy like matching algorithm. When a sub-L-G graph (from the image L-G graph) matches the targets specifications, then that sub-graph is extracted and temporary saved for the next processing task described below. In case that the target is unknown, then the L-G graph of the image frame is exhaustively searched and all possible recognizable or desirable sub-L-G graph forms are extracted and interrelated with their surrounding are saved for further use in later frames.

3.3. TRACKING OF "INTERESTING" MOVING OBJECTS

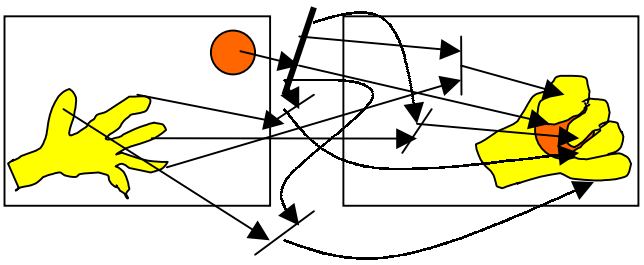
A common approach to tracking non-rigid objects is based on using high-level parametric models representing the various object parts (e.g., legs, arms, trunk, head etc in the case of human tracking) and their connections to each other. However, these methods are difficult to apply in real-world scenes due to the difficulty of acquiring and tracking the requisite model parts (e.g., specific joints such as knees, elbows or ankles in human tracking). Another problem with this approach is that a separate model is required for each type of objects to be tracked (e.g., humans, animals, etc.). In this project, we propose to build dynamic models of appearance of the objects being tracked. These models will enable robust tracking but at the same time provide useful information for classifying moving objects as rigid or non-rigid. The key to building dynamic models of appearance lies on using an incremental eigenspace approach such as the one described earlier for background modeling. Using multiple cameras will facilitate this since different views of the object will be obtained quickly. To build the eigenspace of an object, the segmented moving object will be resampled in a canonical frame throughout tracking, despite changes in scale and position. Its estimated bounding box will be used to resize and resample it into a canonical view. To classify the object as rigid vs non-rigid, the distances of the views used to update the eigenspace will be analyzed (non-rigid objects are expected to create widely different views, thus, the distances will be large compared to the ones obtained when building the eigenspace of rigid objects). To interrelate the detected targets, the system searches for the sub-L-G graph form of the same target. When the sub-L-G graph is discovered , the two graphs are connected by maintaining the relationships between the sub-graphs of the same in consecutive frames . The same process is repeated and the output is the tracking path of a target in a sequence of video frames. When a target is detected in a specific frame, the L-G graph provides the capability of interrelating each recognizable target with any other target in the same frame.

3.4. REPRESENTATION OF ACTIONS AND EVENTS

Definition: An action **Am** represents the mapping from a state (S(i,t) → S(j,t') : **Am** : **Σ**x**S** → **S**, where **Σ** represents the set of actions and **S** the set of states of a target. Thus, an action **Am** could be described as the SPNG equivalent to Gf graph that interrelates the same target into different (or consecutive) frames.

Definition: An event **Ei (Ti,Tj)** between two targets Ti,and Tj is the result of a set of actions **Am** executed by a certain order on these targets.

Example: **Event** (*catching a ball; target-1 the hand, target-2 the ball*), **Actions** (*the five fingers change status from the "open" state into the "closed" state due to coming ball in a synchronized manner described by the SPN*).



An example of a catching event using an SPNG for two consecutive frames: A human hand. and a ball : a) Frame –1: the hand in the open state and a ball in the area coming to its direction; b) Frame-2 : the hand into a closed state catching the ball. The ball plays the role of the triggering action to make the transition of the targets parts (fingers) from one state into another.

3.5. ACTIONS RECOGNITION, EVENTS INTERPRETATION AND ACTIVITY RECOGNITION

The recognition of an action, based on its definition, is based on the efficient representation of the SPNGs associated with the states and transitions involved with the particular target. The recognition process is actually a matching of the states, transitions and their order of appearance against the order of target-states and transitions available in a DB. The interpretation of an event is actually the recognition of the sequence of actions involved in an event. This means that certain sequences of actions will represent events in a DB, thus, if a sequence of actions has been extracted from a video, that sequence goes against the sequences available in the events DB. The human activity is defined as a sequence of actions and a set events. Thus, the recognition of the human activity will be the recognition of the actions and events involved within. Here the system has to be appropriately trained with a variety of activities before its use in real examples.

4. SPECIFIC SCENARIOS WE WILL USE TO DEMONSTRATE OUR METHODS

The methods described here will be used both in controlled and uncontrolled environments. Surveillance in controlled environments, which have specific rules that can be monitored and enforced, has two distinct aspects. The first is surveillance in a highly controlled environment, in which specific access and procedural rules must be followed, for security and safety reasons. These include identification, authorization, and access rules; two-person and other supervision rules; and rules of physical procedure when using sensitive and/or hazardous materials. We also plan to use the surveillance prototype on assist in the monitoring of compliance with the identified rules. The second aspect of controlled environment surveillance is the characterization of general, or even individual, behavior to monitor for anomalous scenarios. In a sensitive facility, this is a way to address the insider threat. Many cases of insider compromise have been characterized by markedly unusual behavior that went unrecognized, from such gross indicators as unusual work schedule changes and after-hours access, to more subtle indicators such as unusual work habits, paths, or access. Although it is not being suggested that every example of anomalous behavior would indicate a significant threat, we know that it is possible, as demonstrated by human observers, to detect certain kinds of unusual behavior that should be alerted. Adaptively learning behavior and calling attention to unusual circumstances for review by the facility monitoring personnel is a complex and very important aspect of this surveillance application. A key application for surveillance in uncontrolled environments, where there are no specific rules for activities to follow, is embodied in monitoring airports for alarming behavior.

5. TESTBED

A good system design is essential to the success of a surveillance system and many computer vision projects ignore this aspect. The testbed for this study is provided by Honeywell. Specifically, Honeywell has already performed preliminary studies and installed a set of cameras in one of their parking lots. Their system design addresses the specification of a camera set arrangement that optimally covers the parking lot [15]. The LANL facilities and data resources will be used in this project.

5.1. An Illustrative Example

In this section we provide an illustrative example to present the potential of the methodology described here. In particular, three consecutive image frames were used, see bellow, to show an activity that takes place after either a verbal command (token), like *"give the cup"* is given, or "*the man is looking for the cup*" by himself. In these synthetic frames, the segmented human is imposed on the original images to visually make clear the sequence of changes (high-level) of the states of his body parts (head, (blue), arm (red), hand (green), legs (yellow)). In particularly, from each image frame the main parts of the segmented human body are extracted, represented and interrelated using the L-G graph. At this point, the SPNG connects these parts in consecutive frames to determine the changes of the states that take place. Between the first and the second frame, the token activates or causes the activation of the head to "look at" a certain direction, at the same time the arm is moving up and the legs change position. The SPNG graphically presents and connects these changes. The same happens between the second and the third frames, where the SPNG connects the body parts using the affect of the previous token. The important issue here is that the SPNG provides the ability of synchronizing the actions performed by each body part. In this particular example the NN model has no visual contribution since the involvement of learning cannot be shown from one example.

Acknowledgement

The authors wish to express many thanks to their colleagues and the supporting agents related with this project.

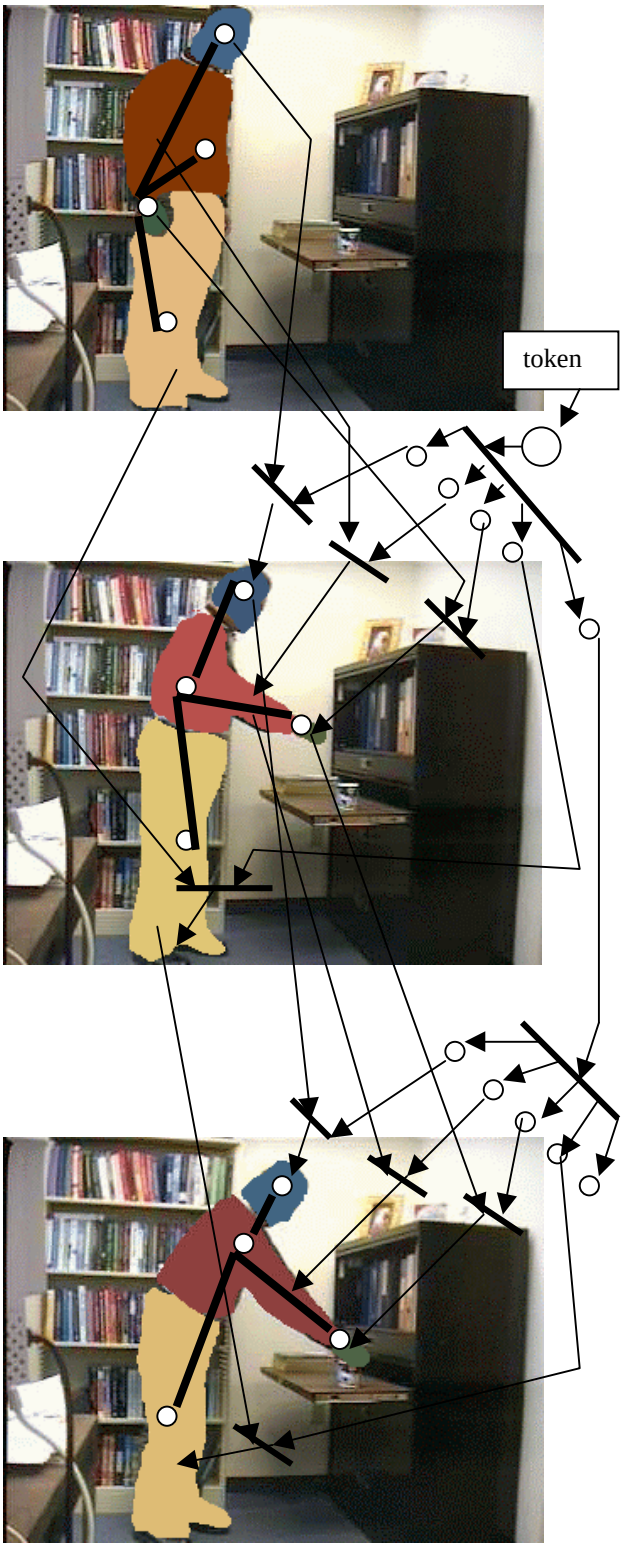


Fig: In this example the L-G graphs and the SPNG graphs are represented by with thick and thin lines respectively

6. REFERENCES

[01] N. Bourbakis, A Neural-based KB using SPNGs in sequence of images, AIRFORCE & SUNY-B-TR-1991,45 pages

[02]J.Gattiker&N.Bourbakis Representation of Structural and Functional Knowledge using SPN Graphs, Proc. IEEE Conf. on SEKE 1995, MD.

[03]N. Bourbakis,HVP: Measuring Human Perception using the L-G graph, SUNY-B-TR- 1997.

[04] N. Bourbakis and R.Andel, ATR from sequence of images Proc. ICTAI, CA,1997

[05] N. Bourbakis, Multimedia Information Systems using SPNGs, ACM Conf. LasVegas, 1997, invited talk.

[06] N. Bourbakis, Facial expressions using L-G&SPNGs , SUNY-B-TR-1997

[07] J.Yamato, J.Ohya and Kenichiro, "Recognizing human action in time-sequential images using Hidden Markov Model", IEEE Conf CV. 1992, pp.379-385

[08] C.Bregler, "Probabilistic Recognition of Human Actions", UCB-TR-May-1996,p.28

[09] C.Bregler, "Learning and Recognizing Human Dynamics in video sequences"IEEE Conf. on CVPR, Puerto Rico, June 1997

[10] J.Feldman and C.Bregler,MICRO-TR-96-103, 1996-97.

[11] S.Bhonsle, A.Gupta, S.Santini, M.Worring and R.Jain, "Complex visual Activity Recognition using a temporally ordered DB", UCSD-TR-1998

[12] A.Madabhushi and JKAggarwal, "Bayesian approach to Human Activity Recognition", UT-Austin, TR-97-ARP-275

[13] M.walter, S.Gong and A.Petrou," Stochastic Temporal Models of Human Activity", UW-TR-UK-1998

[14] R.Rosales and S.Sclaroff, "Trajectory Guided Tracking and Recognition of Actions", BU-CS-TR-99-002, p.29

[15] M.Yeasin and S.Chaudhuri, "Visual understanding of dynamic hand gestures", PR Journal 33,11,2000, 1805-1817.

[16] I.Pavlidis, et. al. "DETER: detection of events for threat evaluation and recognition, to appear in IEEE Proceedings.

[17] G. Bebis, S. Uthiram, and M. Georgio-poulos, " Face Detection and Verification Using Genetic Search", International Journal on Artificial Intelligence Tools, vol 9, no 2, pp. 225-246, 2000.

[18]G. Bebis and K. Fujimura, "An Eigen-space Approach to Eye-Gaze Estimation", ISCA 13th International Conference on Parallel and Distributed Computing Systems, pp. 604-609, Las Vegas, 2000.