Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Image and Signal Processing (CIISP 2007)

Chaos-Synchronization Based Representation of Objects and Events From MPEG-7 Low-Level Descriptors

Hanif Azhar and Aishy Amer
Electrical and Computer Engineering, Concordia University, Montréal, Québec, Canada
{h_azhar, amer}@ece.concordia.ca

Abstract

Chaos theory has been reported to simulate partial functions (i.e., neuronal activity in brain) of the Human Visual System. In this work, we propose a chaotic synchronization-based representation of semantic entities (objects, events) in surveillance scenes using MPEG-7 low low-level Ds. MPEG-7 visual Descriptors(Ds) are used to extract low-level features of video objects. The chaotic synchronization is used to perform feature binding (i.e., group semantically relevant feature elements) from these Ds. The objective is to search for unique numeric descriptions (based on low-level features) to identify semantic entities. The idea of a semantic space is introduced to explain feature binding from multiple features spaces. Subjective evaluation (based on classification) shows the existence of such numeric description for related semantic entities (e.g., male, female, automobile, multiple persons, enter, appear, move).

Keywords: Chaos, MPEG-7, Semantic, Descriptor, Video Surveillance, Classification.

1. Introduction

The research in video surveillance is moving towards semantic (i.e., meaning) analysis of the scene contents through high-level descriptions. High-level description of a surveillance scene identifies semantic entities (i.e., objects, concepts, events) with its meaning. An example of this high-level description is object label. The description also includes concepts (e.g., *indoor, crowd, forest*) or events (e.g., *enter, move, driving car inside forbidden zones*).

Current trends in high-level description generation [1–3] in surveillance video map the pixel-based semantic features of the scene content to low-level MPEG-7 visual Descriptors (Ds) (e.g., DominantColor, RegionShape, MotionActivity, TextureBrowsing). The structural relations among these Ds are then used to identify pre-labeled high-level descriptions.

Latest Human Visual System (HVS) approaches (e.g., [4]) in surveillance video interpretation, focus more on the multiple processing module integration for multiple semantic view of the entities. Multiple view do not confirm the exhibition of intelligence beyond the defining rules of the semantic entities. Some approaches use numerous threshhold dependent rule-based techniques(e.g., [5,6]). All these approaches do not exhibit any additional semantic significance other than that offered by the extracted low-level features.

There have been demonstration of certain aspects of vision in neuro-science experiments [7]. Results in [7] supports the hypothesis that neural trajectories in the human brain are heavily dependent on chaotic activity. Chaotic synchronization plays a significant role in visual information processing [8]. It has been used for image segmentation, chaotic neuron simulation [7] and feature binding [8].

Our work in this paper concentrates on identification of the frame-based objects and events (context-independent interpretation). In the proposed method, MPEG-7 Ds are extracted per video object of interest. Then chaotic synchronization is used to perform feature binding (i.e., group semantically relevant feature elements) of MPEG-7 Ds.

The contribution of this work, is to generate unique numeric description (based on low-level features) to identify a semantic entity. This new description is defined as chaotic descriptor of the corresponding semantic entity in the scene. The chaotic descriptor suppose to exhibits additional semantic significance (supplemental to what is offered by MPEG-7 Ds) of the low-level features.

The rest of this paper is organized as follow. In Sec. 2, we explain how the semantic entities in the surveillance scenes can reside in multiple feature spaces. Sec. 3 describes our proposed approach for chaotic synchronization. The experimental framework is mentioned in Sec. 4. The results and future work is included in Sec. 5 and Sec. 6, respectively.

1-4244-0707-9/07/$25.00 ©2007 IEEE

2. Semantic and Feature Spaces

The semantic classes hierarchy of common semantic entities of interest illustrates that the possible classes in the the feature space is not explicit (i,e., prune to overlapping). For example, it is not trivial to find separate event cluster between *'deposit' and 'remove'* pair or *'enter' and 'appear'* pair. Also due to the limitations of the video analysis (i.e., object segmentation and tracking), the pattern in the data is very un-ruling (e.g., different orientations and postures available for video objects in the same class, partial background inclusions in the bounding box of the video object of interests, poor resolution, multiple objects tracked together as one video objects, multiple events tracked together as one video objects).

We argue that single feature space do not guarantee successful simulation of human perception. The semantic space is a multi-dimensional space (Fig. 1) that is composed of hierarchical semantic layers. Every low-level feature, high-level semantic entities can have a high dimensional vector representation in respective semantic layer (bottom-up hierarchy) in the semantic space. The feature spaces of the MPEG-7 Ds reside in the lower layer of this space (see Fig. 1).

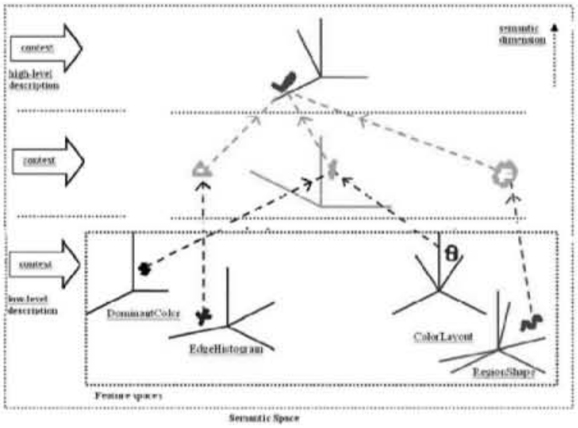

Figure 1. Layered semantic space.

Any semantic entity (e.g., 'enter' event) is composed of the feature vectors in the multiple feature spaces (e.g., Dominant Color, Edge Histogram, Region Shape). The semantic dimension is represented in the y-axis of the Fig. 1 which grows (vertical) according to the semantic significance of the feature vectors. And the x-axis represents the number of low-level features. The dimensions of each feature space is dynamic (vary with the feature vector dimensions per descriptor). Similarly the dimensions of mid and top semantic layer vary with the dimensions of the semantically significant feature vector per semantic entities.

3. Proposed Approach

The basic idea is to consider each element in any MPEG-7 D, as dynamical system with changing behavior in the semantic space of the scene. These dynamical systems are similar to biological *neuron* with numerical descriptions. These assumed neurons can be excited with an appropriate non-linear chaos theory equation (similar to the electro-chemical action potentials of the neurons in the human brain). Chaotic synchronization is defined as the complete coincidence of the trajectories of the coupled individual systems in the trajectory space. Assignment of a dimension value gives a quantitative characterization of the geometrical structure of the chaotic attractor (i.e., chaos equation generated trajectory). We use the algorithm in [9] to compute the Correlation Integral (CI) vector of an individual time series which estimates the embedding dimension of the attractor. The output representation of the chaotic synchronization module is then feed in a classifier to identify objects or events of interest.

In our work, we use two common one-dimensional chaos equations Mackey Glass [9] and Logistic map [9] for first-hand subjective evaluation. In a chaos equation iteration parameter i (of length n) is the time interval that express the changes in the series over that specific dimension (e.g., time i). Instead of considering the iteration parameter i as the time dimesion we consider i as the semantic dimension (see Fig. 1). So in the proposed method a chaotic attractor represents a semantic attractor in the semantic space.

The feature binding in the semantic space in the proposed method is performed as follows. The method outputs unique chaotic descriptor (numeric vectors) for object labels, concepts and events. Then the elements (after normalization) from each extracted Ds are feed in the proposed chaotic synchronization method. Each elements of these Descriptors can be used as seed s, for the chaos time-series equation (e.g., Mackey Glass equation) with a uniform temporal excitation i (e.g. iteration length $n=500$). Thus we can have individual chaotic attractor from each elements of the Ds of the corresponding object. These attractors will have complex geometrical structure of similar pattern. The chaotic synchronization steps are illustrated in Fig. 2, where we also integrate Coupled Map Lattice (CML) [10] for neighborhood interaction of each chaos time series. As we apply the CML, we couple the time series from the Ds. We find CI [9] vector for each coupled time series. This CI vector gives the numeric characteristics for each coupled time series.

To find a scalar representation of each Corelation Integral (CI) vector, we consider either the mean or the median (depending on effectiveness of the corresponding chaotic descriptor) of the CI vector distribution. We identify the largest cluster of the CI vector (from all the chaotic attrac-

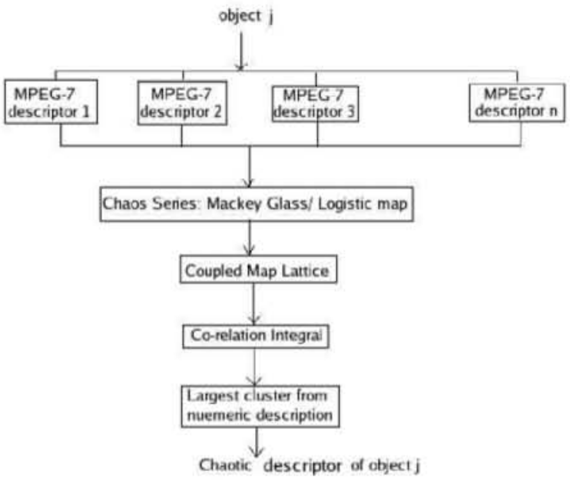

Figure 2. Proposed algorithm block diagram.

tors simulated from the seeds) using histogram analysis. A numeric description can be obtained as output for corresponding semantic entity. We define it as chaotic descriptor (Fig. 2). This signature can be used to label the semantic entity (e.g., male, female, enter, exit) of interest.

4. Proposed Experimental Framework

The proposed experimental framework is given in Fig. 3. The input video scene first goes through video analysis module [5] that produces a list of video objects. Total fifteen (indoor/outdoor) video scenes (i.e., MeetSplit3rdGuy, PET dataset1, cheerleaders, coast, ekrlb, football, intelligent room, road2, road3, survey, urbicande, vand, vlab, vnj, weather). We then select 67 video objects (frame-based) from the above scenes as training data. We use another 198 video objects as testing data. The video objects are prelabeled in different semantic entity classes. Object identification and event (which can be identified from a key frame) are considered separately. The object classes we consider are single person (male, female), multiple persons, trees, boat, bag, phone, automobile, and others. The event classes we consider are enter, exit, appear, move, deposit, remove and others. The 'others' class is used in both the case to label unidentified objects or events.

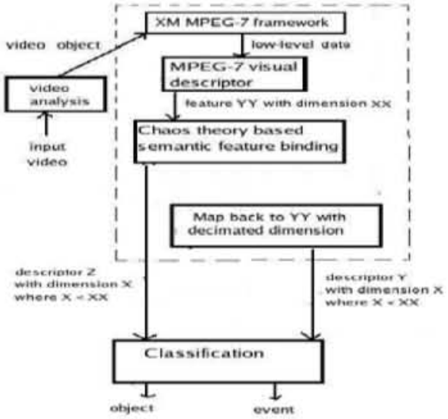
XM MPEG-7 framework
video object
low-level data
MPEG-7 visual descriptor
video analysis
feature YY with dimension XX
input video
Chaos theory based semantic feature binding
Map back to YY with decimated dimension
descriptor Z with dimension X where X < XX
descriptor Y with dimension X where X < XX
Classification
object
event

Figure 3. Experemental framework.

The MPEG-7 reference software XM is used to extract the visual Ds (feature vector YY of dimension XX) of 265 semantic entities (the entities are considered separately for objects and events). Thus from pixel-level features we move to a set of MPEG-7 Ds (i.e., Contour Shape, Region Locater, Region Shape, Color Layout, Dominant Color, Edge Histogram, Scalable Color). An XML parser is used to fetch the values of the elements of the Ds. In total 210 elements are taken from the seven extracted Ds per semantic entity. The use of MPEG-7 Ds in this step is supposed to minimize the first layer of semantic gap between the low-level features and the high-level description (Fig. 1).

5. Results

We report results in this work in three subsections. First we achieve dynamic dimension reduction from the MPEG-7 Descriptors (YY to Y). Second, we present the data quality of YY, Y and Z with statistical coefficents. Third we present unique object and event Descriptors.

5.1. Dynamic Dimension Reduction in YY

The Ds are then used (upon normalization) as seeds to simulate chaos time series (in this case either Mackey Glass or Logistic map). The chaos toolbox from Matlab is used to simulate chaotic series and calculate CI vector. The numeric description from the chaotic synchronization can be defined as Z (with dimension X) per entity. While always X < XX, X does not have fixed dimension per semantic entity. Depending on the chaotic feature-binding output, the dimension of X can vary (number of elements selected from the largest cluster of the CI vector calculated dynamically) per semantic entity. Thus Y exhibits dynamic dimension reduction (different X values for different semantic entities). We have tested our simulation using both mean and median values of the CI vector. The feature elements outside the largest cluster (from the mean of their CI vector values) are replaced with null. The null elements can be discarded as irrelevant for description.

Instead of displaying dynamic dimension X per entity, null is padded to keep the dimension same as XX (as in YY). This padding is done to compare graphical results with YY (of XX dimension). In the simulation with Logistic map, there are total 8672 null elements in YY. There are 12040 null elements in Y. These null elements is scattered (depending on the feature-binding) in various rows of the 168 x210 test video object matrix. Thus the dimension X is not the same for all the 168 video objects. For this reason we claim the dimension in Y is reduced 'dynamically' from XX to X.

The new chaotic numerical representation Z (i.e. YY replacable by Z) has 1267 null values. We called this numerical representation as the chaotic descriptor of the corresponding video objects. Simulation with the median value of CI vector also yield different chaotic descriptor with 2969 null elements in Z. Similarly the simulation with Mackey Glass, there 8672 null elements in YY, there are total 29826 null elements in Y. The new chaotic numerical representation Z has 2028 null values. We called this numerical representation as the chaotic descriptor of the corresponding video objects. Simulation with the median value of CI vector with Mackey glass yield different chaotic descriptor with 2028 null elements in Z.

As more null elements implies more dimension reduction in Y, we prefer to present the chaotic descriptors in this paper from the mean of the CI vector rather than the median.

5.2. Information Gain

A general impression of a data vector can be achieved by calculating the mean, the standard deviation and the entropy of the vector. The mean shows the average location of the underlying feature extraction method. The standard deviation expresses idea on the discriminance. If the standard deviation is near zero, any feature extraction yield same output for any type of given content. A higher entropy of a data vector reports more interesting and variant information than a lower entropy of other vector.

The mean, standard deviation, and entropy for YY, Y and Z are shown in Fig. 4 (from Logistic map simulation) and in Fig. 5 (Mackey-Glass equation) for all the 198 test video objects. We see that with Logistic map, Z gives higher entropy than that of Y and YY. The mean and standard deviation of Z is close but lower than Y and YY. The values lie between 0.0 to 0.4. With Mackey Glass equation, the entropy of Z is higher than Y but less than YY. The mean and standard deviation is less than that of Y and YY. As simulations with Logistic map give higher entropy, in the rest of this paper we represent results from Logistic map simulation only.

We have accomplished two new numeric vectors (Y and Z) of reduced dimensions to describe objects and events of interest other than MPEG-7 Descriptors (YY). When YY has dimension XX both Y and Z has dimension X (X < XX). While the mean, standard deviation and entropy of Z is not always better than YY, for Y (with reduced dimension X) the values are similar to (if not better) than YY.

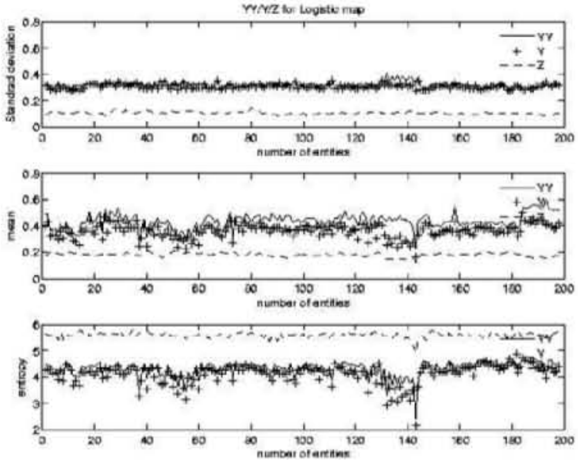

Figure 4. Standard deviation, Mean, and Entropy of 198 x 210 test video object matrix (Logistic map simulation).

5.3. Identification from Classification

To see the classification performance on the test video objects we apply two different classifiers (5-Nearest Neighbor and Genetic Algorithm) for both object classes and event classes (separately for YY, Y and Z to observe). 5-Nearest Neighbor is chosen as one of the simple classifier. As our data is more complicated and of high-dimension models, we get slightly better identification from the output of stochastic classification methods (e.g., Genetic Algorithm). The collection of the training and testing video objects and pre-labeling them to generate the ground truth is done manually. First the video analysis module is applied on the fifteen video scenes. Only few *'video objects of interest'* is selected from the thousands of frames. Our ground truth generation yields very few number of training data for some of the object and event classes.

These few training data are inadequate for successful classification analysis. Yet we look for the unique chaos descriptors that represents the semantic entities successfully even with poor classification. Same classifier is used for all of the YY, Y and Z. We report the cases where the test video objects are successfully identified (compared to the ground truth) by the Z from corresponding chaotic descriptors. In some cases the video objects are misclassified by YY, yet classified correctly by Z (in some cases also by Y).

We show the vector Z using Logistic map equation for female and automobile in Fig. 6, and Fig. 7 respectively. We only show three of the test video objects which are classified successfully (i.e., matched with ground truth). The chaotic descriptors are the mean values of the CI vector. In all the

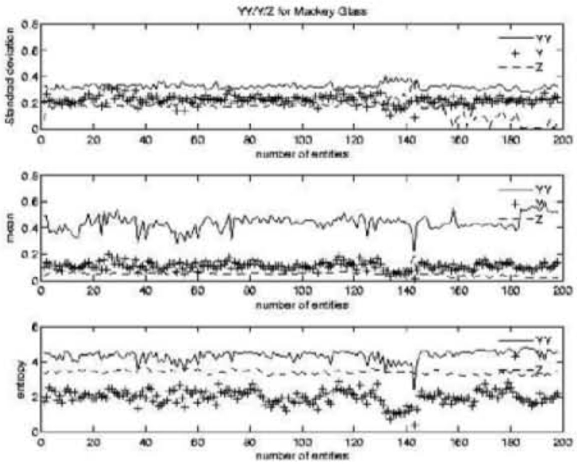

Figure 5. Standard deviation, Mean, and Entropy of 198 x 210 test video object matrix (Mackey Glass simulation).

above case we see unique pattern in the description of male, female and automobile. Similarly the chaotic descriptor for different frame-based events (e.g., remove and deposit) are presented in Fig. 8 and Fig. 9 respectively. Selecting an appropriate chaos time series that successfully simulate the semantic relations in surveillance video scenes, is a major challenge. Other related chaos equations needs to be investigated in future simulations for this purpose.

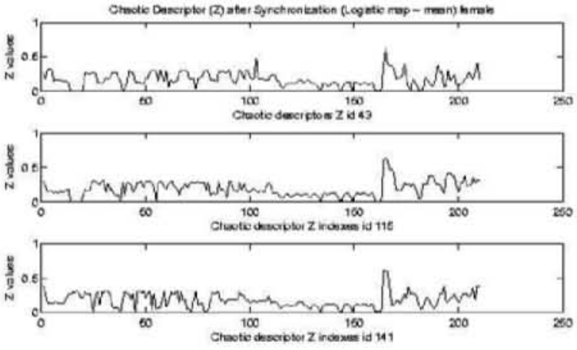

Figure 6. Z with Logistic map (female)

In Sec. 5.2, it is noted that both the Y and Z provide either very similar information gain. So we graphically compare the description Y (reduced dimension from YY) for different objects and events with that of YY and Z. The comparison with corresponding unique numeric description is shown for multiple person in Fig. 10. Apart from using Z, the numeric description Y (dynamically reduced dimension from YY) also can be used to replace YY (MPEG-7 Descriptors) for object or event identification.

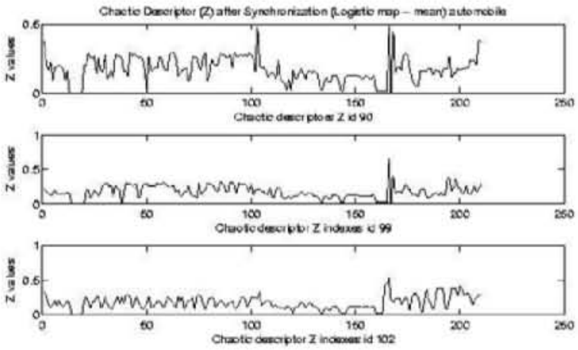

Figure 7. Z with Logistic map (automobile)

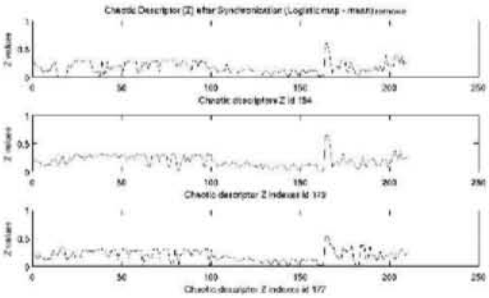

Figure 8. Z with Logistic map (remove event)

6. Conclusion and Future Work

In this paper, chaos synchronization is performed to find out the non-linear dependencies among different MPEG-7 Ds elements of the semantic entities in any frame-based scene representation. The effectiveness of the output chaotic descriptor is related to the set of low-level features used in the simulation. For different set of Ds elements the corresponding chaotic descriptor will vary.

Combination of the semantic description strength of the MPEG-7 standard with that of chaotic synchronization in the interpretation of surveillance video is a new approach. The focus in video processing has been on the structural (i.e., geometrical and statistical) relations (mainly linear) of the features. Chaos theory has not been used in video processing significantly other than video compression. Nonleaner relations and non-structural semantic relations are, however, getting more importance in video processing (e.g., in video surveillance) research. Our work shows that the aperiodic nonlinear characteristics in chaos theory can be utilized to identify semantic entities.

Currently the performance evaluation is mainly depen-

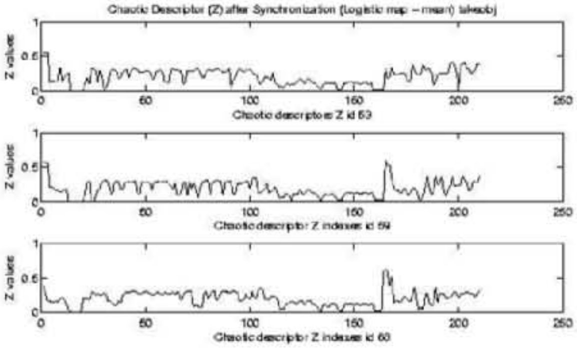

Figure 9. Z with Logistic map (deposit event)

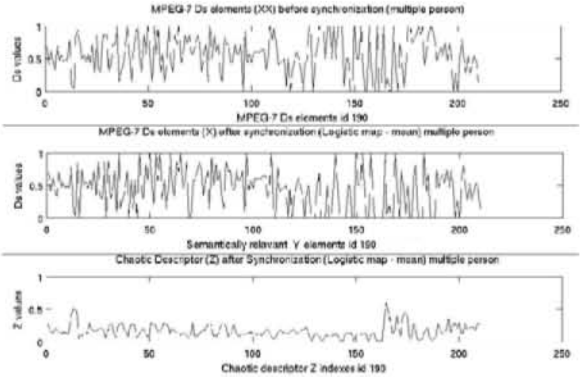

Figure 10. YY, Y, Z with Logistic map (multiple person)

dent on subjective observation of the chaotic descriptor. In this paper we also apply two classification techniques (k-Nearest Neighbor and Genetic Algorithm) on the test video objects with all three representations mentioned in this report(i.e., YY, Y and Z). While we achieved limited classification accuracy maninly due to 1) the un-ruling pattern in the tracked video objects (Sec. 2) and ((Sec. 5.3)) limited number of ground truth for training and test video objects.

Successful object or event classification is not the objective of our work. We accomplish unique feature based chaotic descriptor to identify specific objects and events of interest as comprehensive as possible from the corresponding surveillance scene.

Our future work includes the addition of more MPEG-7 Ds to further strengthen the chaotic feature binding to yield more distinctive chaotic descriptor for semantic entities. Also more suitable ground truth (for both training and testing) will be selected to significantly make the chaotic descriptor more suitable for classification. Multiple frame-based Ds elements of the video objects (e.g., motion activity, trajectory) in the proposed method also will be integrated for multiple frame-based semantic entities.

References

[1] O. Steiger, "Adaptive Video Delivery Using Semantics," Ph.D. dissertation, Swiss Federal Institute of Technology (EPFL), Lausanne, 2005.

[2] L. Xin and T. Tan, "Ontology-Based Hierarchical Conceptual Model for Semantic Representation of Events in Dynamic Scenes," in *2nd Joint IEEE Int.Workshop on VS-PETS, Beijing*, Oct 2005, pp. 57–64.

[3] I. Lin, "Video Object Plane Extraction and Representation: Theory and Application," Ph.D. dissertation, Electrical Engineering, Princeton University, USA, 2000.

[4] T. List, J. Bins, R. B. Fisher, and D. Tweed, "A Plug-and-Play Architecture for Cognitive Video Stream Analysis," in *IEEE Int. Workshop on Computer Architecture for Machine Perception*, Palermo, Italy, 2005, pp. 67–72.

[5] A. Amer, E. Dubois, and A. Mitiche, "Real-time System for High-level Video Representation: Application to Video Surveillance," *Elsevier Journal for Real-Time Imaging*, vol. 11, no. 3, pp. 244–256, 2005.

[6] H. Li, "Hierarchical Video Semantic Annotation-The Vision and Technique," Ph.D. dissertation, Electrical Engineering, The Ohio State University, USA, 2003.

[7] C. A. Skarda and W. J. Freeman, "How Brains Make Chaos in Order to Make Sense of the World," *Behav Brain Sci.*, vol. 10, pp. 161–195, 1987.

[8] F. T. Arecchi, "Chaotic Neuron Dynamics, Synchronization and Feature Binding," *Computational Neuroscience: Cortical Dynamics Lecture Notes in Computer Science*, vol. 3146, pp. 90–108, 2004.

[9] P. Grassberger and I. Procaccia, "Characterization of Strange Attractors," *Phys. Rev. Lett.*, vol. 50, pp. 346–349, 1983.

[10] K. Kaneko, "Spatiotemporal Chaos in One- and Two-Dimensional Coupled Map Lattices," *Physica*, vol. 37 D, pp. 60–82, 1989.