# Summarizing Video Datasets in the Spatiotemporal Domain

A. Stefanidis, P. Partsinevelos, P. Agouris, P. Doucette
*Dept. of Spatial Information Engineering*
*National Center for Geographic Information and Analysis*
*University of Maine*
*5711 Boardman Hall #348*
*Orono, ME 04469-5711*
*USA*
*{tony, panos, peggy, doucette}@spatial.maine.edu*

## Abstract

*In this paper we address the problem of analyzing and managing complex dynamic scenes captured in video. We present an approach to summarize video datasets by analyzing the trajectories of objects within them. Our work is based on the identification of nodes in these trajectories as critical points that describe the behavior of an object over a video segment. The time instances that correspond to these nodes are used to select critical frames for a video summary that describes adequately and concisely an object's behavior within a video segment. The analysis of relative positions of objects of interest within the video feed may dictate the selection of additional critical frames, to ensure the separability of converging trajectories. The paper presents a framework for video summarization using this approach, and addresses the use of self-organizing maps to identify trajectory nodes.*

## 1. Introduction

The aspect of time becomes increasingly important in modern geospatial applications. In addition to traditional discrete multitemporal datasets (e.g. maps of the same area at various time instances), dynamic events are also captured in video datasets. Video data processing and analysis, and video database management present well-known challenges, mostly associated with the size and complexity of the information space that has to be explored.

In this paper, we address the problem of analyzing and managing complex dynamic scenes depicted in video datasets. Efficient modeling of dynamic environments is an important step towards the analysis and management of large video datasets. In modern geospatial applications, dynamic environments tend to be multidimensional, including spatial and temporal dimensions complemented by content and knowledge. The objective of the process described in this paper is the generation of information-rich summaries of video scenes that will describe the spatio-temporal behavior of objects within a depicted scene. These summaries are envisioned as new concise multimedia videos comprising vectors and images that portray the significant parts of the original video dataset.

The development of concise representation schemes is essential for the search, retrieval, interchange, query, and visualization of the information included in video datasets. Efforts towards this direction include attempts to summarize video by selecting discrete frames at standard temporal intervals (e.g. every n seconds). However, such an approach would typically fail to capture and represent the actual content of the original video dataset. Summarization alternatives include the use of image templates, statistical features and histogram based retrieval and processing [1]. Video summaries have also been proposed, taking into consideration both visual and speech properties to construct a "skim" video that represents a synopsis of the original video. This "skim" Video is constructed by merging segments of the original video [2]. Video posters are proposed alternatives to describe story content [3], while [4] has presented approaches to identify different scenes within a video stream by analyzing a variety of properties (e.g. dominant motion, and various histogram properties). In the trajectory domain, for fixed environments, systems extract and recognize moving objects, and classify their motion. [5], [6]. In addition, generation of spatiotemporal synthetic datasets to simulate movement trajectories is seen in [7].

Here we present an approach to summarize video datasets by analyzing the trajectories of objects within

them. Our work is based on the identification of nodes in these trajectories as critical points in the video stream. These nodes form a generalization of the trajectory of a moving object within a video stream. The time instances that correspond to these nodes provide the critical frames for a video summary that describes adequately and concisely an object's behavior within a video segment. In doing so, we benefit from substantial advancements in object extraction from digital imagery, and video image processing.

The paper presents a framework for video summarization using this approach. Section 2 offers an overview of our overall approach. In section 3 we present the role of nodes in generalization, together with an algorithm for their selection. Section 4 shows the extension of this concept for multiple objects, and section 5 presents the use of this summarization in hierarchical data models. Experimental results are used throughout the paper to demonstrate the performance of the designed algorithms.

## 2. Overview of Proposed Approach

In monitoring applications, the background usually remains fixed while objects move throughout the scene (e.g. cars moving in a parking lot monitored by a camera atop a nearby building). In such an environment, the crucial elements for video generalization are those describing the behaviors of the moving objects in time. We consider the spatiotemporal space of a scene as comprising two *(x,y)* spatial dimensions and one *(t)* temporal dimension. Object movements are identified by tracing objects in this 3-dimensional (x, y, t) space. These trajectories are the basic elements upon which our summarization scheme is based. An outline of our approach is shown in Fig. 1. The spatial coordinates are those defined by the image space*(x,y)*. If we want to translate them into survey coordinates we can use any of the well-known orientation models and relevant pose/rotation parameters of the specific video camera.

We base our selection of representative frames on the segmentation of trajectory lines into break points termed "nodes". The nodes are distributed dynamically to capture the information content of regions within the above mentioned 3-D S-T space. More nodes are assigned where trajectory presents S-T breakpoints (e.g. moving at fixed velocity in a straight line from point A to point B), and fewer nodes are assigned to segments where the spatiotemporal behavior of an object is smooth. Node placement is based on concepts of self organizing maps (SOM) from neural network theory. The number of nodes may be selected to control the degree of generalization, similar to the number of nodes in a k-means approach. Using more nodes (resp. fewer) will result in lower (resp. higher) generalization of the original video signal. The spatiotemporal trajectories and generalization nodes lay down the framework to express the content of video datasets via tree-like hierarchical data structures. These hierarchical data models describe object movement in a scene, and are based on the expansion of standard octrees.

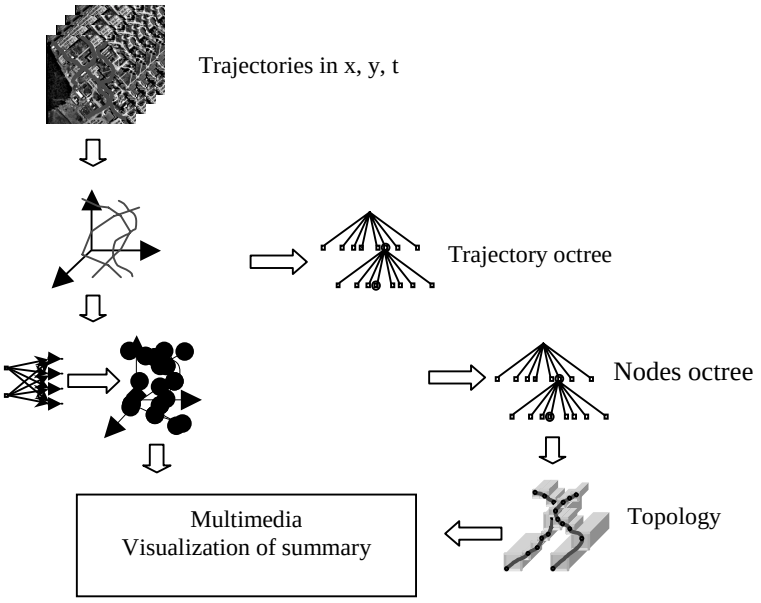


Figure 1. Outline of the proposed approach

The summary of a video is a new shorter video, which includes a base map-image representing the background of the monitored area. Actual video frames are used at node instances, while the behavior of objects between nodes is represented by rapidly evolving vectors (e.g. moving spots or trace lines). The time it takes to bridge nodes in a video summary depends on the desired duration of the video summary. It is analogous to selecting a tape speed to fast forward between events. High speed results into shorter, more concise video summaries. The choice of generalization resolution is a function of the application at hand and specific user needs. This visualization aspect is beyond the focus of this paper; we focus on node selection in this scenario.

Combined, the above provide a novel approach to manage dynamic scene analysis at higher levels of abstraction, and to visualize concisely the behavior of moving relations in a scene. In doing so we benefit from advancements in computer vision and digital image analysis, and transfer methodologies from these areas into spatiotemporal analysis. The above outlined processes offer a robust and consistent way to describe the content of video datasets, and provide a powerful environment for further analysis. This new abstract environment of data summaries is a first step toward complex scene understanding, behavior comparisons, and information dissemination. While we assume a fixed sensor, we could easily handle the case of a moving sensor by relating the variable video coordinates to the fixed axes of a suitable mosaic model [8].

## 3. Summarization of Video Data

### 3.1. Generalization in S-T Cubes

Temporal generalization in a video sequence is equivalent to dividing the time coordinate in varying intervals. In our approach we do not address the generalization of the spatial coordinates (x, y) as methods like scale space analysis offer readily solution to that problem. These intervals can be defined dynamically by incorporating spatial or other knowledge. According to the desired degree of generalization, the intervals may increase (higher generalization) or decrease (lower generalization). Our main consideration with a video sequence is to capture the behavior of moving objects and their relations within a fixed-background scene. Therefore, our generalization procedure is primarily driven by these moving objects.

In the previously defined S-T cube, each frame is registered at the time *(t)* of its acquisition. Assuming a fixed camera, the spatial dimensions *(x,y)* of the cube coincide with the image coordinate systems of each individual frame. In that sense, individual frames pile up on top of each other to form the 3-D S-T cube. The movement of an object within this cube manifests as a set of point clouds (e.g. resulting from an image classification) moving over time. Treating the S-T cube as a quasi-continuous representation of reality, the trajectory of an object defines a linear feature within this 3-D space, by connecting all positions of the same object over time. The trajectory begins at point $(x^i_0, y^i_0, t^i_0)$ and ends at point $(x^i_n, y^i_n, t^i_n)$, where $(x^i_0, y^i_0)$ are the image coordinates of object i at the time $t^i_0$ that it first appears in the video field of view, and $(x^i_n, y^i_n)$ are the corresponding coordinates at the time $t^i_n$ that it moves outside the video field of view. Generalization of the video stream will be based on the generalization of this trajectory. Trajectory generalization will be performed by distributing nodes over this linear feature. The spacing of nodes is performed automatically, using SOM technique to capture the information content of the given trajectory.

### 3.2 Single Object Trajectory Analysis using SOM

The self-organizing map (SOM) algorithm [9,10] is a nonlinear and nonparametric regression solution to a class of vector quantization problems, which is used herein as the method for information abstraction. The SOM belongs to a distinct class of artificial neural networks (ANN) characterized by unsupervised and competitive learning. Essentially an iterative clustering technique, the SOM differs from traditional clustering methods, such as *K*-means or ISODATA, in three fundamental ways: 1) cluster centers are spatially *ordered* in the input space according to a predefined topology, 2) a shrinking neighborhood function is used to act as a smoothing kernel over the cluster center adjustments, and 3) cluster centers are updated sequentially versus batch. The *network space* $\Re_N$ exists independent of the input space $\Re_I$, and the objective of the SOM is to define a mapping from $\Re_I^m$ onto $\Re_N^d$ where $m \geq d$. To demonstrate, let $p(\mathbf{X})$ describe a probability density function in $\Re_I^2$ for the input vector,

$$\mathbf{X} = [x, y]^T \in \Re_I^2, \quad (1)$$

Each connection or *synapse*, between a component of **X** and any single node *k* located in network space has an associated weight. The components of each weight vector are defined in $\Re_I^2$, which has the same dimensionality of **X**, or,

$$\mathbf{W}_k = [w_{k,x}, w_{k,y}]^T \in \Re_I^2 . \quad (2)$$

By initializing the contents of **W** in $\Re_I^2$ for each node, the goal of competitive learning is to reward the node *k* that optimally satisfies a similarity measure between a given **X**

compared against all $\mathbf{W}_k$ . Using the $L_2$ (Euclidean) norm as the similarity metric, a *winning* node $q$ is determined as,

$$\text{node}\, q = \arg\min_k \| \mathbf{X} - \mathbf{W}_k \|, \ \text{for}\, k = 1, 2, ..., K. \quad (3)$$

where $K$ is the total number of nodes in $\Re_N^1$. The appropriate weight vectors are updated sequentially for each input sample according to Kohonen's learning rule,

$$\mathbf{W}_k(n+1) = \mathbf{W}_k(n) \ + \ \eta(t) \cdot h_q(t) \cdot \left(\mathbf{X}(n) - \mathbf{W}_k(n)\right) \quad (4)$$

Here, $\mathbf{X}(n)$ represents the $n$-th sample drawn from $N$ total input space samples, $\mathbf{W}_k(n)$ are the node weights at the $n$-th iteration, and $\mathbf{W}_k(n+1)$ are the updated weights for the $n$-th iteration. A time variable $t$ is measured in epochs, each of which represents a complete presentation of $N$ input samples to the network. A learning rate function, defined as $0 < \eta(t) < 1$, dynamically controls the relative rate of weight updates. The neighborhood function $h_q(t)$ centered on the winning node, is defined as $0 < h_q(t) \leq 1$. The network nodes adapt to the local density fluctuations in $p(\mathbf{X})$ through *ordering* and *refinement* phases, during which $h_q(t) \to 0$ as $t \to \infty$. Multiple epochs (iterations) are typically required for asymptotic convergence of the algorithm. The basic SOM algorithm is summarized as follows [11]:

1. Initialize the synaptic weight vectors $\mathbf{W}(n=1)$ for $K$ nodes.
2. Randomly draw an unseen sample $\mathbf{X}(n)$ from the input space.
3. Determine the winning node $q$ using a similarity metric as in eq. (3).
4. Update $\mathbf{W}$ for winners using eq. (4).
5. Return to step 2, and iterate until stopping criteria (checked after each epoch) are satisfied.

SOM is used in [12], as a technique in which robust road delineation within a noisy image environment is performed. Extending the SOM to the ST domain is illustrated in fig.2, in which an 8-node neural chain is used to abstract local velocity fluctuations from a moving object.

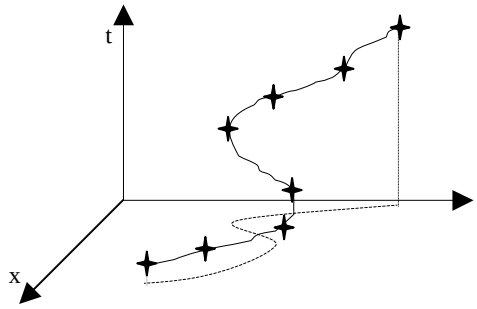


Figure 2. Information abstraction in the ST domain.

## 4. Multiple object trajectory analysis

The consideration of multiple objects brings forward the need to address two issues. First, we have to select specific time instances for our video summaries using independent nodes from multiple trajectories. Second, we have to consider the introduction of additional nodes when taking into account the proximity of two or more trajectories. One can easily understand that the set of temporal coordinates of the nodes describing the path of object i and those describing the path of another object j may be totally disjoint. According to the density and the dissimilarity of the S-T trajectories and the corresponding nodes, we can follow different strategies for merging a complex scene summary:

An obvious solution is to use the nodes from all S-T trajectories and reference all moving objects to every estimated node. This results in a relatively large summary, depending on the number and behavior of the objects.

Another solution is to define nodes according to the most demanding moving object and project all other node sets to this dominant set. If the behaviors of scene objects are incompatible then the other objects are not efficiently represented. One way to overcome the overcrowded node collection is to group sets of nodes over a minimal increment $\delta t$ and identify an average temporal position $t_{av}$ to substitute individual nodes. This allows us to minimize the number of nodes and the complexity of the produced summary.

Furthermore, by using the SOM we can obtain a "medium" estimation of node selection upon the whole set of moving objects. This gives a summary of the whole scene, which does not explicitly depict behavior information for single objects. On the other hand, it provides a technique to unravel mass behavioral attitudes in the scene. For instance if a police car enters the scene, the majority of the moving cars tend to slow down.

The relations between two or more moving objects define significant information. Therefore, we introduce mandatory nodes termed as "relational" when the proximity defined in the S-T cube between two or more objects drops below a threshold. Furthermore, we can introduce spatial regions of interest within which we wish to monitor more closely the behavior of objects (e.g. the entrance of a bank within the broader field of view of a monitoring video). In the S-T cube we construct new 3-dimensional regions and when a trajectory passes through these regions, we add mandatory nodes in order to include this information in the summary. These types of nodes are termed as "reasoning" nodes. In fig.3 we show the S-T trajectories of two objects. The area of interest is projected as a cylinder in the S-T space and it defines two reasoning nodes while a relational node is introduced

when the proximity between the two trajectories reach the threshold.

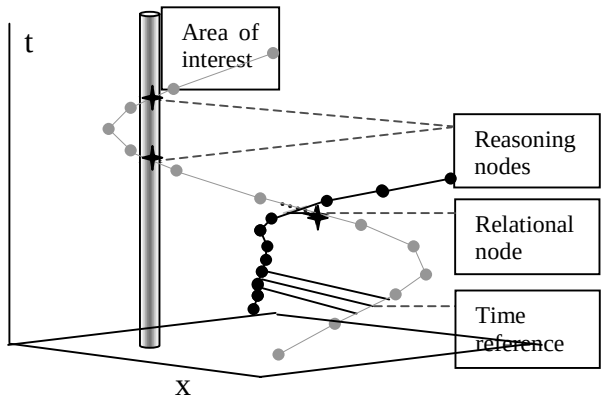


Figure 3. Mandatory nodes for multiple S-T trajectories.

## 5. Hierarchical data models

In the previous sections, we presented a methodology to transfer video content from the original video to the S-T trajectory space. We now present the modeling of trajectories by representing them with hierarchical tree structures. Use of hierarchical data structures provides higher level of information and leads to computationally less expensive management of large datasets. Sorting data according to their spatial occupancy through tree structures is a promising data manipulation scheme [13, 14]. Topological spatial relations support spatial analysis with focus on relations in a higher information level where further processing is accommodated.

According to the octree structure, decomposition of data volume is performed iteratively in a step by step fashion by dividing the space into eight disjoint cubes with the aim of eventually meeting a resolution criterion. If any of the cells is homogenous, i.e. the cell lies entirely inside or outside an object, or satisfies the resolution criterion, the sub-division stops. If the cell is heterogeneous then it is sub-divided further into eight sub-cells until the prespecified criteria are met. The information on the original video stream is compactly represented and the leaf nodes represent minimum resolution segments. In the S-T trajectory space, the process produces a summary of the information while in the node domain we model the S-T occupancy of the nodes.

### 5.1. S-T trajectory space

The octree deals with the representation of a continuous curve or a group of discrete yet numerous points. The characteristics of the octree design are the following:

- Input data is a 3D set of points (x, y, t)i.
- The decomposition process is based on standard octree decomposition as described above.
- The termination criterion is of dual nature. First, it depends on the complexity of the curve included in the cube. If the curve is not complex there is no need for further decomposition. If a complexity threshold is met then we continue decomposition of S-T space. Second, if the cube reaches a predefined volume value then the process terminates even when the complexity of the curve is still large. Complexity is defined by measuring the angle of the intersection of the cube with the curve. Another more precise measure of the complexity is by taking the summation of the measured angle between every three points of the curve and project it to the distance between them. The decomposition process, criterion checking and the form of the octree is shown in figure 4.

### 5.2. S-T node space

In the node space the design is similar and it includes:

- The input data is again a set of points of (x, y, t) space but they are sparse.
- Decomposition of S-T space remains the same.
- The subdivision is terminated when each cube has at most one node.

Accordingly, the decomposition procedure and the termination criterion is shown in the fig.5 while the tree is of the same structure. In the octree domain codes and algorithms are available for further processing and for recognizing efficiently some particular pattern in the image, aiding behavior matching and querying in a coarser information level.

## 6. Concluding Remarks

The figures presented in the previous sections are early implementation results from our work in the research direction of video summarization. They demonstrate the capability of the presented approach to generate information-rich summaries of video datasets to convey the spatiotemporal behavior of moving objects within them. The use of digital image analysis techniques for this summarization offers great advantages, the potential for full automation, and the ability for objective yet meaningful analysis of video datasets with being but two among many such advantages.

One could argue that the greatest advantage of the presented approach lies in the potential offered for subsequent analysis. The S-T trajectories of moving objects and their summarization enable the subsequent

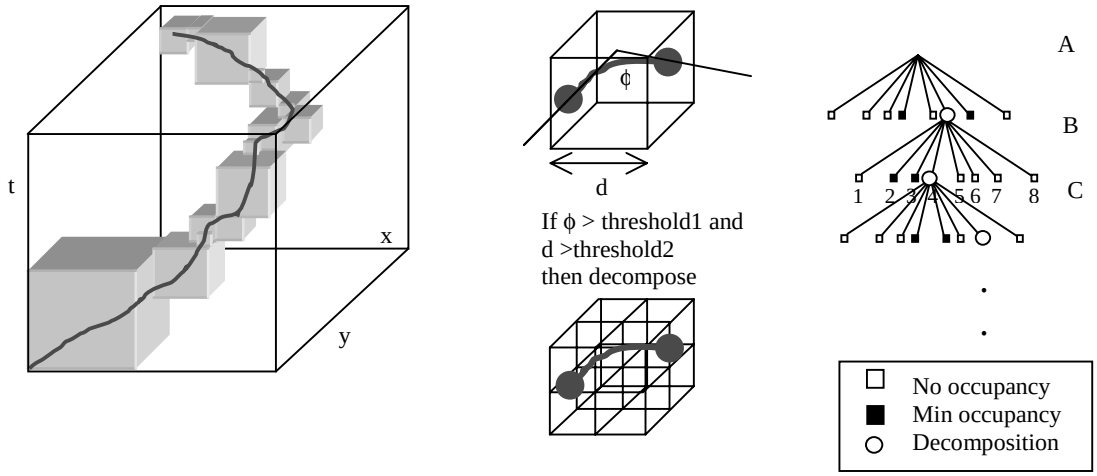


Figure 4. Decomposition of space and winning cubes, criterion check and Tree structure.

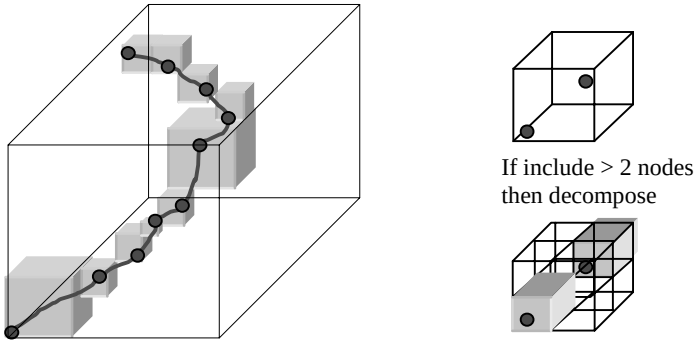


Figure 5. Decomposition of space, winning cubes and criterion check of S-T node octree

analysis of complex behavioral patterns. By treating object movements as linear features we can take advantage of numerous digital image analysis techniques (especially image matching) to establish similarities among such movements. These comparisons can lead to the establishment of behavioral trends, and the identification of complex behavioral patterns. Our future work will concentrate on the establishment of metrics for such analysis.

## 7. Acknowledgments

This work was supported by the National Science Foundation through CAREER grant number IIS-9702233 and Digital Government award number DGI-9983445, and by the National Imagery and Mapping Agency through NURI grant number NMA202-98-1-1113. In addition, Panayotis Partsinevelos is partly supported by the State Scholarship Foundation of Greece, and Peter Doucette by the National Aeronautics and Space

Administration through NASA grant fellowship number MSTF 99-59.

## 8. References

[1] W. Chang, G. Sheikholeslami, J. Wang, and A. Zhang, “Data resource selection in distributed visual information systems”, *IEEE Transactions on Knowledge and Data Engineering*, vol.10, (no.6), , Nov/Dec 1998, pp.926-946.

[2] M. Smith, and T. Kanade, “Video Skimming for Quick Browsing based on Audio and Image Characterization”, *Tech. report CMU-CS-95-186*, Computer Science Department, Carnegie Mellon University, July 1995.

[3] M. Yeung, and Boon-Lock Yeo, “Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content”, *IEEE Trans. on Circuits and Systems for Video Technology*, Vol.7, NO. 5, October 1997, pp. 771-785.

[4] N. Vasconcelos, and A. Lippman, "A Spatiotemporal Motion Model for Video Summarization”, *CVPR*, Santa Barbara, 1998.

[5] G. Medioni, R. Nevatia, and I. Cohen, “Event Detection and Analysis from Video Streams”, *DARPA98*, 1998, pp63-72.

[6] R. Rosales, and S. Sclaroff, “3D Trajectory for Tracking Multiple Objects and Trajectory Guided Recognition of Actions”, *CVPR*, June 1999.

[7] D. Pfoser, and Y. Theodoridis, “Generating Semantics-Based Trajectories of Moving Objects”, *International Workshop on Emerging Technologies for Geo-Based Applications*, Ascona, Switzerland, 2000.

[8] G. Zhou, J. Albertz, and K. Gwinner, “Extracting 3D Information Using Spatio-Temporal Analysis of Aerial Image Sequences” *PE&RS*, Vol. 65, No. 7, July 1999, pp. 823-832.

[9] Kohonen, T., *Self-Organizing Maps*, Springer-Verlag, 1997.

[10] T. Kohonen, “Self-organized formation of topologically correct feature maps”, *Biological Cybernetics*, 1982, pp. 59-69.

[11] Haykin, S., 1999. *Neural Networks*. Upper Saddle River, New Jersey, Prentice Hall.

[12] P. Doucette, P. Agouris, M. Musavi, and A. Stefanidis, “Automated Extraction of Linear Features from Aerial Imagery using Kohonen Learning and GIS”, *Lecture Notes in Computer Science Vol. 1737*, 1999.

[13] Samet H, The *Design and Analysis of Spatial Data Structures* Addison-Wesley, Reading, MA, 1990.

[14] T. Sellis, N. Roussopoulos, and C. Faloutsos, “The R+-tree: A dynamic index for multi-dimensional objects”, *Proceedings of the thirteenth International Conference on VLDB '87*, Brighton, England, 1987, pp. 507-518.