CIHSPS 2005 – IEEE International Conference on
Computational Intelligence for Homeland Security and Personal Safety
Orlando, FL, USA, 31 March – 1 April 2005

Development of an Artificial Intelligence System for Detection and Visualization of Auto Theft Recovery Patterns

Olcay Kursun[1], Kenneth Reynolds[2], Ronald Eaglin[1], Bing Chen[1], Michael Georgiopoulos[3]

[1]Department of Engineering Technology
[2]Department of Criminal Justice and Legal Studies
[3]Department of Electrical and Computer Engineering
University of Central Florida, Orlando, FL 32816
Phone: (407)-823-5937 Fax: (407)-823-5483
{okursun, kreynold, reaglin, bchen, michaelg}@mail.ucf.edu

Abstract – *Auto theft is the most expensive property crime that is on the rise across the nation. The prediction of auto drop-off locations can increase the probability of offender apprehension. For successful prediction, first the patterns of thefts are identified. Then, a prototype expert system successfully identified embedded drop-off location clusters that were previously unknown to investigators. The system was developed using the expert knowledge of auto theft investigators along with spatial and temporal auto theft event data. Drop-off clusters were identified and validated. A map interface allows the user to visualize the feature clusters and produce detailed reports. Such GIS applications give us the ability to attain a geographical perspective of incidents within the community, thus help law enforcement officers discover the patterns of incidents and take necessary measures to prevent them.*

Keywords – *Auto theft and recovery, clustering, crime mapping, data sharing, GIS, hot-spot analysis, pattern discovery.*

I. INTRODUCTION

According to a recent statistics, every 20 seconds a motor vehicle is stolen in the United States. Law enforcement officers work daily to locate and retrieve stolen cars, while insurance companies spend billions of dollars each year compensating owners of stolen vehicles. Auto theft incidents may have complex and varied patterns due to very different objectives such as for auto parts, joy rides, and use in other crimes; thus, computer aided crime analysis tools are required to cop with this crime that is on the rise nationwide.

Hot-spots are places that have shown persistent tendencies to be sites of crime [1, 2, 3, 4, 5, 6]. Discovery of hot-spots is a critical task because the deployment of police and other prevention resources at these hot-spots makes the greatest contribution to crime. Cluster analysis can be an effective method for determining areas with high concentrations of crime [7, 8]. For auto theft events, since the number of potential targets is large, another alternative to prevent auto theft is to capture the criminals at the place of the drop by finding hot drop-locations.

However, it is a particularly challenging task to detect hot-spots by clustering analysis due to the uncertainty of the appropriate number of clusters to generate and the significance of the found clusters [8]. In this work, while avoiding these two big challenges of cluster analysis for hot-spot detection, we have managed to develop a scalable tactical crime analysis tool for auto theft events, specifically for identifying and predicting drop-locations through a cluster analysis approach. In this work, the analyst does not have to specify how many clusters to be found; instead, the analyst is supposed to define what makes a cluster a set of related events. For example, auto theft detectives we have met in Orange County Sheriff's Office agree on the fact that each thief has preferred types, makes, models and years of vehicles to steal and drop-locations based on the means of transportation and where he/she lives/works and where he/she may feel comfortable dropping the stolen vehicles, etc… Although, the process of deciding what should make a cluster a real cluster may seem to be a detailed and involved one; in reality, the analyst always knows what he/she is looking for in the data.

II. AUTO-THEFT DATA

The dataset we used for our preliminary simulations have approximately a thousand auto theft events. This is a small fraction of all the events we have in our database because the rest of the entries do not have a matched address that could be converted into numerical (X, Y) coordinates. The X and Y for the recovery addresses are generated by a geo-coding server. In our experiments, for X and Y coordinates, we used UTM (Universal Transverse Mercator) coordinates using ARCIMS as the geo-coding server. Each event is characterized by the following five features: Make of the vehicle, Year of the vehicle, X and Y coordinates of the recovery location, Date of the theft.

Make of the stolen vehicle is a unique numerical value for each different make. In our dataset, we had 62 different makes of vehicles. Year of the vehicle is a numerical value ranging from 1970 up to date. The (X, Y) coordinates are UTM coordinates (floating point numbers).

III. PATTERN DISCOVERY

To be able to understand the underlying dynamics and patterns of criminal activity, a visualization tool is needed. In order to combat the complexity of the crime spot maps, due to the increasing number of spots, we developed an animation tool that visualizes the criminal activity on a daily basis. This animation tool creates a movie of the criminal activity, where each frame contains auto thefts committed in a day. Such a tool has been used to analyze the spatial and temporal dimensions of gun recoveries by the Bureau of Alcohol, Tobacco and Firearms and shown useful [9].

However, the amount of daily activity can still be too complex to display in one frame because the analyst might not be able to make sense out of the high number of spots being displayed. Therefore, we developed an algorithm to cluster events that have similar patterns, thus breaking down the complexity of the set of all events by partitioning this set into clusters. The commonalities that our algorithm identifies are the kind of commonalities that auto theft detectives would look for in these events. Consequently, our algorithm automates the laborious manual process that these detectives have to go through.

To measure how similar two given auto theft events are we define a distance measure between them. This distance measure is based on a weighted Mahalanobis distance measure (Eq. 1), where the weights, *w*, for each feature *i*, are determined by the domain expert.

$$Dist_{X,Y} = \sum_i \left(w_i \cdot \frac{\left| X_i - Y_i \right|^2}{\sigma_i^{\,2}} \right), \qquad (1)$$

In (Eq. 1), X_i and Y_i represent the value of the i^{th} feature of the incident *X* and *Y*, respectively. The standard deviation of the feature *i* is denoted by σ_i. Since the weights are only relative, it is not challenging to come up with a suitable assignment for them for the task to be accomplished. For example, if the spatial proximity is much more important than temporal proximity, then the weights of X and Y must also be much higher than the weight of the Date of the theft.

The basic idea behind a clustering algorithm is to group together the events with a high degree of similarity amongst them (small inter-distances). The null hypothesis is that all the auto theft incidents are independent of each other (the distribution is perfectly random). Our approach is closely related to the nearest neighbor measure [8], where each new incident is placed in an existing cluster, if it is the nearest cluster to the incident and the distance from the incident to the cluster is smaller than or equal to a predefined threshold value. Otherwise, the incident is placed in a newly created cluster. The threshold is a simple function of the average distance of all the data points as shown in Eq. 2.

$$Threshold = AverageDistance \cdot Sensitivity \qquad (2)$$

The sensitivity is the desired deviation from the null hypothesis. When the sensitivity is set to zero, only repeat-crime activities (exact same pattern and location) are clustered. If it is set to an extremely large value, most of the incidents will be considered to be related. Our experiments showed that the sensitivity should be kept below 0.5 to get feasible clusters (it can be even smaller based on how compact clusters are desired).

Determination of the value the average distance of all the data points does not necessarily require having all the data points in hand. In fact, it only requires some reasonable number of data points to initially estimate the AverageDistance, which is to be re-estimated as more data points are entered to keep it up-to-date.

Thus, our approach does not require a priory specification of the number of clusters to be found. Instead of asking the user-define how many clusters to be found, which brings significant subjectivity into the analysis; we take advantage of knowledge and experience of auto theft detectives to decide the criteria for clustering. Furthermore, there are no established methods in the statistics literature to determine the appropriate number of clusters [8].

To increase the robustness of our clustering algorithm, we also use an upper-bound for the differences of features if two events are to be placed in the same cluster. To illustrate this, consider two events that are a year apart but very similar in other aspects. With a proper selection of the upper-bound for the difference of the dates of the events, we prevent our clustering algorithm from assigning these two events into the same cluster. Such an upper-bound is necessary for a tactical crime analysis, in order to guarantee that two events so far apart in time are indeed grouped into different clusters. However, if desired, the upper-bound can be discarded and the weights can be adjusted to fight such criminal activities. For example, the two events mentioned above can be identified to be in the same cluster even if they were 10 years apart by setting the weight of time to zero and removing the upper-bound. A brief pseudo-code of our clustering algorithm is given below.

For each new entry *N*,

i. **For** all clusters, find D_i: the distance of *N* to the i^{th} cluster (Eq. 1)

ii. **Set** *D* equal to D_m: the minimum of D_i

iii. **If** *D* is not greater than the threshold in (Eq. 2) then *N* belongs to the cluster *m*

Else, a new cluster *C* is created and *N* is placed into the cluster *C*

IV. MAPPING AND VISUALIZATION

We use Oracle 8i to store and manage the auto theft data, each record of which is associated with a spatial column computed from X/Y coordinates. Each X/Y coordinate pair is geocoded from the address field of each auto theft record with ArcIMS, an ESRI GIS product that is also used to visualize and publish the auto theft data online. Another ESRI product ArcSDE, A Spatial Data Engine, is used as a gateway to access this data by ArcIMS. Web Server (Microsoft IIS 4.0), ArcIMS and Oracle form a three-tier architecture that eases data visualization and publishing. Whenever users operate on the map (for example, when a data point on the map is selected, or a query tool in the toolbar is used, or the map is zoomed in/out), the browser (the client) will send the request to the Web server, which then forwards the request to the ArcIMS Server. The ArcIMS Server queries the data stored in Oracle database via ArcSDE. Once the query is done, ArcIMS sends the result (generally a JPEG image) to the Web Server, which then sends the JPEG image to the browser.

On the map interface, three relatively big clusters are shown in Figure 1. For convenience, each event is assigned a short string (A, B, C, etc) by a simple hashing function so that the user can observe the theft and drop/recovery locations of vehicles. A red circle corresponds to a theft location and a blue rectangle corresponds to a recovery location. Using the tools on the left panel, the user is able zoom in/out, select data points, run queries to display only certain points, measure the distance between data points, etc…

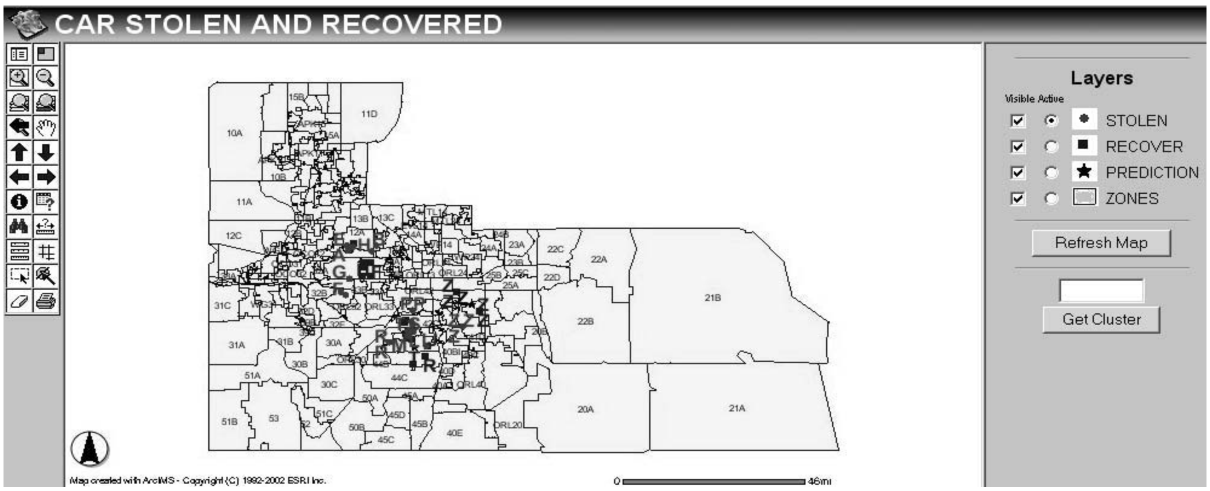

Fig. 1. The map interface for data visualization.

In Figure 2, we demonstrate the use of the interface by zooming in on two exemplary auto theft events identified to be in the same cluster by our algorithm. The user can be automatically informed on the existence of such suspiciously similar patterns of criminal activity with a click of a button. Thus, using this interface with our algorithm allows the police officers easily decide which areas to patrol. Any auto theft even near points N and M (shown with circles) is very likely to be dropped somewhere near points N and M (shown with rectangles).

V. SIMULATION RESULTS

The weights and the upper-limits of the features for the calculation of the distance measure are shown in Table 1. As expected, the clusters found by the algorithm show orderly features such as specific drop-locations for a specific make and years of stolen vehicles (see Table 2). When viewed by our animation tool, each cluster is a collection of auto thefts local in time and recovery location. Breaking down a set of events by such sequences demonstrate spatio-temporal patterns that are often more predictably reliable than those discovered by static statistical methods [10] because the local patterns may disappear when looking at the global statistics.

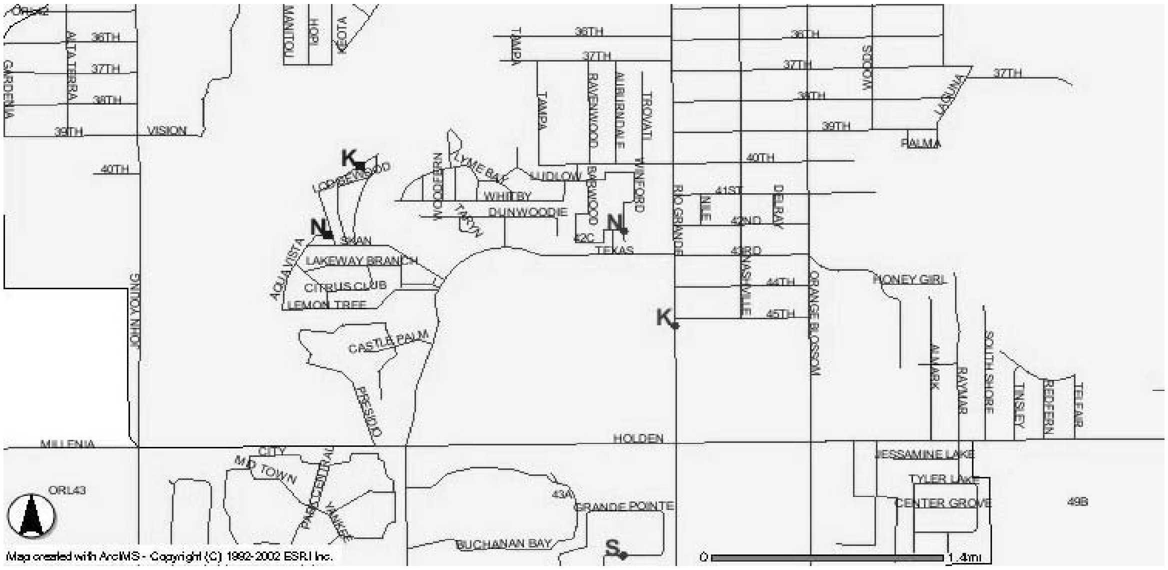

Fig. 2. Zoom in on two exemplary incidents identified to be in the same cluster.

Feature ID	Description	Weight	Upper-bound
1	**Make** of the vehicle	20.0	Not used
2	**Year** of the vehicle	2.5	10 years
3	**X-coordinate** of the recovery location	15.0	Not used
4	**Y-coordinate** of the recovery location	15.0	Not used
5	**Date** of the theft	10.0	60 days

Table 1. The features, weights, and upper-bounds in the simulation dataset. The values of the weights and the upper-bounds are chosen based on empirically the suggestions of the Orange county Sheriff's Department detectives.

As a verification of the quality of our results, we also show the model of the stolen vehicles in the last column of Table 2. Although the model information is not used in the simulations (the algorithm has no hint of the model of the vehicle), some of the clusters are composed of not only the same make but also the same model of the vehicle. In addition to the five fields available in the dataset, every record in a cluster is given an ID number to ease references to specific records.

X-coor	Y-coor	YEAR	MAKE	DATE	MODEL*
504071	1540191	1994	HONDA	10/10/02	ACCORD
506029	1543959	1996	HONDA	10/11/02	ACCORD
501894	1544772	1997	HONDA	10/12/02	CIVIC
508456	1540932	1996	HONDA	10/15/02	ACCORD
510511	1541151	1994	HONDA	10/18/02	ACCORD
511384	1541098	1995	HONDA	10/19/02	ACCORD
510519	1541301	1997	HONDA	10/20/02	ACCORD
506034	1544827	1996	HONDA	10/23/02	ACCORD
511622	1542343	1994	HONDA	10/24/02	ACCORD
513009	1540875	1994	HONDA	10/25/02	ACCORD
510841	1543096	1997	HONDA	11/01/02	ACCORD
507674	1539989	1994	HONDA	11/04/02	ACCORD
503693	1540456	1994	HONDA	11/05/02	ACCORD
511984	1542603	1994	HONDA	11/08/02	ACCORD
514597	1541363	1994	HONDA	11/08/02	ACCORD
506053	1535721	1994	HONDA	11/08/02	ACCORD
511551	1540004	1995	HONDA	11/08/02	ACCORD
504396	1543033	1996	HONDA	11/08/02	ACCORD
514597	1541363	1994	HONDA	11/09/02	ACCORD
513492	1535572	1995	HONDA	11/15/02	ACCORD
513543	1540593	1997	HONDA	12/04/02	ACCORD

Table 2. A representative example of the clusters found by the algorithm.

As seen in Table 2, this cluster consists of Honda Accords, with the record number three as the only exception. Our further experiments show that with a more strict similarity measure (greater deviation from average), we can purify the clusters. In that case, the above cluster will have only the "Accords"; however, the total number of clusters will increase. Therefore, this is a decision to be made based on the specifications dictated by the analyst.

VI. CONCLUSIONS

In this work, to assist law enforcement officers in auto-theft crime analysis, we first developed an animated map that can simulate a given set of criminal activity as point stimuli on a computer spot map. Then, we developed a clustering algorithm to report events that show similar patterns (over time and space) on the map [5, 6, 11]. When animated, the clustered events show hotspot-like, orderly characteristics that are very likely to be the reflections of the underlying patterns of the criminal events.

Even though developing GIS applications can be time consuming and costly, it is worth having the ability to attain a geographical perspective of incidents within the community [12]. GIS also eases the visualization of data, which helps law enforcement officers discover the patterns of incidents and take necessary measures to prevent them [10, 12].

However, it should be kept in mind that small-scale GIS applications are far from efficient in monitoring criminal activity unless they are made compatible for data-sharing. The main reason is that criminals do not respect jurisdictional borders and stolen vehicles in one jurisdiction can be dropped in some other jurisdiction. Such incidents are either hard to follow because of difficulty of obtaining relevant data from other databases, or they cause duplicate efforts for all the neighboring jurisdictions. Therefore, the criminal justice agencies must coordinate with other agencies to arrange for data sharing as they develop new automation systems.

Department of Criminal Justice and Department of Engineering Technology at the University of Central Florida have collaborated with Orange County Sheriff's Office and most of the Florida counties to eliminate duplicate efforts and create opportunities to suppress cross-jurisdictional criminal activity [6, 13]. In preventing auto-thefts, one must consider that a small number of known criminals are responsible for a big chunk of auto theft incidents, the police officers should be aware of bail/parole/release information on these offenders, which can be easily accomplished by such data sharing abilities.

ACKNOWLEDGEMENT

The authors wish to acknowledge helpful discussions with Kunal Motwani and Balaji Chandrasekaran from the University of Central Florida, and Chief Ernie Scott and the other personnel of auto-theft division from Orange County Sheriff's Office.

REFERENCES

[1] Goldsmith, V., McGuire, P.G., Mollenkopf, J.B., Ross, T.A. (1999) *Analyzing Crime Patterns: Frontiers of Practice.* Sage, California.

[2] Hick, S.R., Bair, S., Fritz, N.J., Helms, D. (2004) Crime Mapping. In: Bruce, C.W., Hick, S.R., Cooper, J.P. (eds) *Exploring Crime Analysis*, IACA Press, Overland Park, KS.

[3] Levine, N. (2000) CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations. *Crime Mapping News.* Vol: 2(1), pp. 8-9.

[4] Mena, J. (2004) *Homeland Security Techniques and Technologies.* Butterworth-Heinemann.

[5] Reynolds, K., Kursun, O., Georgiopoulos, M., Eaglin, R. (2004) Development of an Artificial Intelligence Clustering Algorithm to Detect Auto theft Recovery Patterns. GIS Symposium 2004, Troy State University, Troy, AL.

[6] Reynolds, K., Scott, E., Eaglin, R., Pan, P., Kursun, O. (2005) A GIS Based Artificial Intelligence Clustering Algorithm to Detect Auto-Theft Recovery Patterns. *The 3rd UK National Crime Mapping Conference,* London, UK.

[7] Pattavina, A. (2005) Geographic Information Systems and Crime Mapping in Criminal Justice Agencies. In: Pattavina, A. (ed) *Information Technology and the Criminal Justice System*, Sage Publications, Thousand Oaks, CA.

[8] Grubesic, T.H., Murray, A.T. (2001) Detecting Hot Spots Using Cluster Analysis and GIS. *The Fifth Annual International Crime Mapping Research Conference: Translating Spatial Research into Practice,* Texas.

[9] Freeman, J.R. (2002) Extending GIS to Visualize Investigative Data: Some Practical Examples from the ATF. *The Sixth Annual International Crime Mapping Research Conference: Bridging the Gap Between Research and Practice,* Colorado.

[10] Helms, D. (2002) Criminal Movement Pattern Detection. *The Sixth Annual International Crime Mapping Research Conference: Bridging the Gap Between Research and Practice,* Colorado.

[11] Cooper, S. (2004, December 7) Innovative Program Identifies Auto-Theft Patterns. *University of Central Florida College of Health and Public Affairs Newsletter. http://www.cohpa.ucf.edu/news/story_auto-theft.cfm.*

[12] Hughes, K. (2000) Implementing a GIS Application: Lessons Learned in a Law Enforcement Environment. *Crime Mapping News.* Vol: 2(1), pp. 1-5.

[13] Gutierrez, P.R. (2005, February 6) Database connects cops. *Orlando Sentinel.*