# Can Human Visual Surveillance be Improved with Intent Recognition?

**Alireza TAVAKKOLI**
**Department of Computer Science, University of Houston-Victoria**
**Victoria, TX 77901, USA**

**and**

**Donald LOFFREDO**

**Department of Psychology, University of Houston-Victoria**
**Victoria, TX 77901, USA**

## ABSTRACT

In video surveillance applications, trained operators watch a number of screens simultaneously to detect potential security threats. Looking for such events in real time, in multiple videos simultaneously, is cognitively challenging for human operators. This study suggests that there is a significant need to use an automated video analysis system to aid human perception of security events in video surveillance applications. In this paper the performance of humans in observing a simulated environment is studied and quantified. Furthermore, this paper proposes an automated mechanism to detect events before they occur by means of an automated intent recognition system. Upon the detection of a potential event the proposed mechanism communicates the location of such potential threat to the human operator to redirect attention to the areas of interest within the video. Studying the improvements achieved by applying the intent recognition into the simulated video surveillance application in a two phase trial supports the need for an automated event detection approach in improving human video surveillance performance. Moreover, this paper presents a comparison of the performance in video surveillance with and without the aid of the intent recognition mechanism.

**Keywords:** Object Tracking, Intent Recognition, Activity Detection, Video Surveillance, Analysis of Variance.

## 1. INTRODUCTION

Maintaining the security of public and private buildings is an essential priority for governmental agencies, private industries, and individual home owners. The smallest commercial video surveillance packages setup four video cameras and feed the videos on a 2x2 grid of display screens. To cover large areas for video surveillance more cameras are required. This increases the number of cameras and screening monitors to 9 or even more. Performing video surveillance tasks in real time by monitoring the screens is a very challenging process. Trained security personnel have to constantly watch several screens to detect potential security threats from simultaneous video feeds of secured areas. The detection of threats becomes cognitively more challenging if multiple events occur almost simultaneously in different videos and shown on different screens.

Kasturi et al. in [1], studied the effectiveness of human operators in performing video surveillance tasks on 2x2 and 3x3 grids. Since 3x3 and even 4x4 monitor grids are common in video surveillance applications, their study showed a considerable need for augmenting video surveillance tasks to improve their efficiency. However, the study used simple graphics -i.e. 2 dimensional rectangles as the screens and moving boxes as people. The study concluded that under this simplified graphical model the performance of volunteers in detecting single, double, and triple events suffered as the tasks became more complicated.

In this paper, we propose an automated mechanism based on object tracking [2], and intent recognition [3], to assist human operators in detecting security events in video surveillance applications. In our study a factory-like simulated environment is designed in an advanced 3D game engine to produce video images for the surveillance application. By creating a simulated factory-like environment, the performance of human operators are expected to be close to real video surveillance applications since the videos are physically realistic.

One important aspect in modifying visual surveillance tasks is to reduce the perceptual and cognitive demands on human operators. To this end, the proposed automated system processes images from multiple sources -i.e. security cameras- to detect and track moving objects of interest. The tracking trajectories are used by the intent recognition module for the purpose of predicting possible events before they occur. The screen(s) which may contain events are highlighted by the automated mechanism to redirect the observer's attention and achieve a higher change of catch-

ing such potential threats.

Section 2, below, provides the background about the computational tools employed for processing video surveillance applications and the review of the literature. Section 3, discusses our methodology for implementing the object tracking and the intent recognition modules. In section 4, detailed quantitative and statistical results of our system are presented and a discussion about the expected and achieved outcome of the study is given. Section 5 concludes the study and provides future possible directions of the work.

## 2. LITERATURE REVIEW

The detection of the intent of people from videos has been a recent study area within the field of computer vision [4]. To develop reliable and efficient intent recognition there are two main tasks to perform. The first process looks at detecting humans from a stream of video images [5]. The main objective of object tracking is to provide tracking trajectories for objects of interest in the scene [6].

These trajectories are employed in a computational framework to calculate probabilities of future possible activities [7]. These activities, before they occur, may carry information about the intent of individuals participating in the action. Evaluating and predicting the intent of agents present in the video is the main objective of object tracking mechanisms.

In the following an overview of the literature on object tracking and intent recognition approaches is presented.

### Object Tracking

The process of object tracking requires a low-level modeling of the background as well as the appearance of the objects of interest. Most of the state-of-the-art foreground detection algorithms model the color appearance of the background pixels statistically either non-parametrically [6], or by specifying the parameters of the density functions governing the distribution of the pixel colors [8] or variations of these two main tracks [9].

The advantage of the parametric modeling techniques is their low memory requirements. However, the non-parametric techniques require that the distribution of the background pixel colors to be known, or at least assumed, heuristically. On the other hand, the non-parametric background modeling techniques do not require the pixel color distribution be known apriori. Unfortunately, for the model to be trained accurately the non-parametric modeling approaches have large memory requirements. To alleviate these problems, the proposed framework employs an incremental modeling approach based on Support Vector Data Description [10].

Once foreground objects are detected their geometric or photometric appearances could be used for tracking purposes. For rigid objects, whose geometric appearance does not undergo major changes, a data association mechanism such as Kalman filtering [11] is able to provide sufficient tracking accuracy. However, more sophisticated tools such as Monte Carlo Sequential Importance Re-sampling [12] employed in Particle Filters [13], and Hidden Markov Models [14] may be used. But these techniques suffer from a great computational cost. To address the speed of the probabilistic models based on Monte Carlo Re-sampling, Comaniciu et al. in [15] proposed a histogram-based tracking approach based on non-parametric appearance modeling.

Most recent tracking approaches use an attention-based localization method [16] and an interconnected target detection/tracking loop [17] for object tracking. A significant issue with these current tracking algorithms is their slow speed as well as poor scalability in tracking a large number of objects. In this paper a tracking model is proposed based on finding correspondences between detected objects and their photometric appearance of known objects.

### Intent Recognition

For a robotic or an intelligent agent to successfully communicate with humans, it is very critical to understand the potential intents of humans with whom it is interacting. Although natural to humans, even in their early developments, endowing intelligent agents with such capabilities has proven to be difficult. The general principle on which the intent recognition may be formalized relies on the psychological evidence of the Theory of Mind [18]. The premise is that humans understand about others' intentions by taking their perspectives [19] while using their own experiences to infer about the potential intents [20].

The intent recognition system approaches the lower level intentions similar to the models proposed in [3]. However, the models for the passing-by and the meeting intentions are combined to achieve a higher detection rate. In this visual surveillance application, the passing-by scenario loses its meaning while the intention to meet can be modeled as the intention to enter a forbidden area, when such contextual clues exist. This also differs from the modeled intent proposed by Gray et al. in [21], in that the intent to enter a forbidden area conveys lower level task goals with the integration of the context.

## 3. METHODOLOGY

The objective of this work is to implement a computational framework for simulating visual surveillance applications and to employ the simulated scenarios in an automated mechanism to detect potential security events. As discussed, simulated videos are very challenging to human observers since they contain different moving objects with different movement patterns. Moreover, the simulation is physically realistic and represents a factory-like environment. Therefore, human observers will be faced with several perceptual challenges to detect events as well as the difficulty of simultaneously watching multiple screens.

This architecture serves as the essential platform for quantitatively studying the efficiency of human visual

surveillance. An advanced 3D game engine, Unreal Development Kit (UDK), is used to create the simulated environment with multiple moving robots and human guards. Videos of several scenarios are shot within this environment from different camera location. The system is employed for evaluating performance improvement obtained within visual surveillance tasks, by the proposed processes for the recognition of animated objects' intents.

Participants were given six videos to watch and to interact with the computer to record the time-line of the security events they detected. Three of the six videos simulated a 9 camera surveillance system showing simulated videos taken from 9 different locations in a 3x3 grid. The other three videos simulated a 4 camera surveillance system.

Videos recorded from the cameras in the simulated environment were either without an event or contained one of three types of events. Singular events were defined as a single security breach. A single security breach occurs if an agent enters a forbidden area in one of the 4 or 9 videos in the case of 2x2 and 3x3 surveillance systems, respectively. Double events contained two security threats occurring almost simultaneously in two different videos in 2x2 or 3x3 surveillance scenarios. Finally, triple events occur when three almost simultaneous security breaches are present in three of the 9 or 4 videos.

A two phase trial is conducted to study the statistical significance of the proposed intent recognition models. 54 undergraduate college students were recruited to participate in the study. All volunteers participated in both phases of the study.

In the first phase of the study the participants watched raw videos and were asked to detect the event. Upon finding a single, double or triple event, the participants were asked to record them by interacting with the evaluator program. The timeline of all events detected or missed by each participant was recorded to study their performance in detecting each event type without the aid of an automated intent recognition mechanism.

The participants were asked to return for the second phase of the study two weeks after the first phase. In the second phase, participants observed similar videos processed with the intent recognition module. In this case the intent recognition highlighted one, two or three screens out of 4 or 9 simulated monitors, when there was a high probability of security breach(s).

Once the two phases of the study were complete, a comprehensive statistical analysis based on Multivariate Analysis of Variance (MANOVA) was performed to find the statistical significance of the intent recognition in improving the efficiency of human observers. In particular we were interested in achieving a system-independent performance between the 2x2 and 3x3 simulated system settings in the detection of single, double or triple events.

**The Object Tracking Mechanism**

The object tracking mechanism uses the foreground regions detected by a background segmentation mechanism. The background segmentation step of the proposed architecture models the boundary of the color distribution of each pixel in the video frames. This boundary is an analytical description of the distribution and as such is not bound to the statistical accuracy of the probability estimation methods. The analytical boundaries of the pixel colors are trained using an Incremental Support Vector Data Description [10].

```
1. For each video frame
1.1. Receive Foreground Masks
1.2. Detect Connected Components (CC)
1.3. Process CCs to detect blobs.
1.4. Process Collision Potential
1.4.1. If No Collision
         Build Spectral Model
         Connected Component Spectral Tracker
         Model Update
1.4.2. If Collision
         Stop occluded object(s) model update
         Connected Component Spectral Tracker
         for the occluding object
           Meanshift tracking for occluded objects
1.4.3. Kalman update and post process
```

Figure 1: The object tracking algorithm.

Once the foreground regions are detected relevant regions are processes to detect contiguous blobs for the object of interest, i.e. moving objects. The spatial proximity of connected components detected from the background segmentation step are employed to detect each object of interest. Figure 1 shows the pseudo-algorithm of the proposed object tracking mechanism. The novel Spatio-Spectral Connected Component (SSPCC) tracking mechanism performed in step 1.4.1 carries the core tracking module of the proposed algorithm.

The SSPCC tracker maintains a list object whose photometric models are created and maintained. The photometric model is generated from the RGB color value of pixels in video frames. The object list is originally empty before any object is introduced into the scene. In every frame the tracker maintains a correspondence matching process to assign the list of unlabeled blobs with an appropriate photometric model produced so far.

The tracker also maintains the geometric information of each blob for the purpose of collision detection. Among geometric features we use objects' center of gravity, width, and height. These geometric moments, as well as the tracking trajectories for each pair of blobs, are used in predicting their possible location in the next frame to detect the possibility of a collision.

The photometric models employed in the SSPCC tracker are composed of two components. The upper model maintains the first order statistical model of colors of all pixels in the upper half of the object, while the lower model represents the first order statistical representation of the lower part of the object's pixel color. The models are shown in the following equation where $O_u$ is the model for the upper part and $O_l$ is the model for the lower part of the object $i$ at

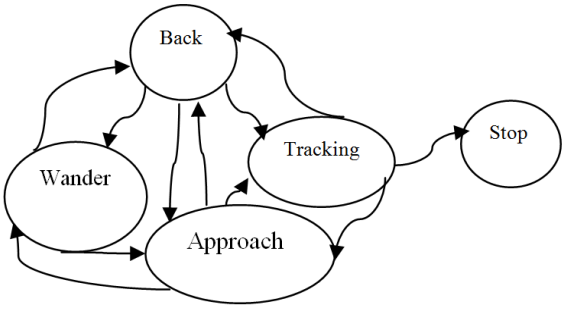



Figure 2: The intent recognition Hidden Markov Model.

time $t$.

$$O_{u,l}^{t} = \mathcal{N}\left(\mu_i^t, \sigma_i^t\right) \quad : \quad \forall i \tag{1}$$

In the process of correspondence matching each blob is divided into an upper and a lower part. The first order statistics of the upper and lower parts are calculated and matched with each model encountered and maintained by the SSPCC. The model with the highest matching score will be used to label and track the blob. If the matching scores are low, a new model for the blob is created and maintained for the future. The correspondence matching assigns the label of the highest scoring model to the blob whose statistical models closely resemble the model. This is shown in the following:

$$\forall k: \quad O_{u,l}^{t-1}(j) = arg\max_{k} \left[ median \begin{pmatrix} P\left(c_l^t(j)|O_l^{t-1}(k)\right) \\ \times P\left(c_u^t(j)|O_u^{t-1}(k)\right) \end{pmatrix} \right] \tag{2}$$

where $C$ is the current blob being matched with $O$, the model from the previous frame.

**The Intent Recognition Module**

In the proposed approach a method for constructing the model is chosen in formulating models for an agent's interaction with the world while performing the activity. This is done through the way in which parameters that encode the goals of the task are changing (e.g. increase, decrease, and stay constant or unknown). This is in contrast with the traditional approaches that solely model transitions between static states. With this representation, the visible states encode the changes in task goal parameters and the hidden states represent the hidden underlying intent of the performed actions [4].

The reason for choosing the activity goals as the parameters that are monitored by the HMM is that goals carry intentional meanings, and thus tracking their evolution is essential for detecting and understanding an agent's intent.

In the proposed intent recognition mechanism a model is developed based on each observed moving object in the scene and a location designated as the forbidden area. The parameters of the moving object with respect to the designated forbidden area are used as observable variables and a probability for each observable state is calculated. The Hidden Markov Model and its corresponding intentional hidden states are shown in the Figure 2.

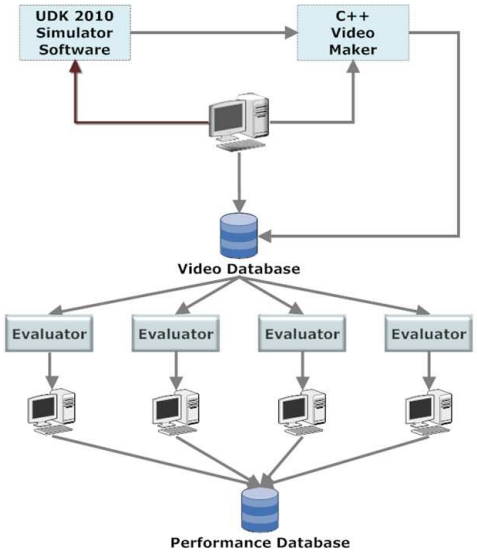



Figure 3: The proposed visual surveillance evaluation architecture.

**Simulated Environment and Experiment Setup**

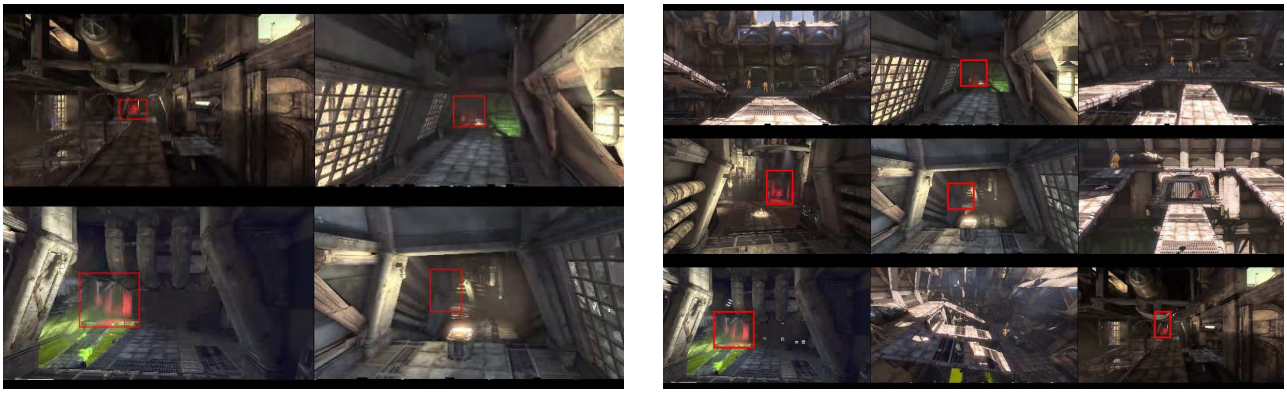
The computational framework for this project has been developed with various advanced computer graphics, multimedia and programming software packages. The complete implementation of the proposed video surveillance architecture is shown in the diagram in Figure 3.

The primary means for implementation of the platform to simulate the realistic graphics of visual surveillance scenarios is based on the Unreal Development Kit 2010 game engine. The simulated scenarios are rendered in two programs implemented with the C++ platform to develop test videos and to evaluate the performance of the subjects.

Once the visual surveillance scenarios are created within the UDK, the frames are used in a rendering program developed as a part of this project. In order to carry out the rendering tasks the module uses the results obtained from the intent recognition component to highlight the respective videos.

The rendering program generates the final 2x2 and 3x3 grid sizes for each visual surveillance scenario. Realistic animated objects - i.e. robots and people - wander in the scene. Some simulated scenarios may contain one or more violations of a number of secured regions resulting in "events". The performance of human subjects in detecting one of more events accurately will be quantified.

Figure 4 shows a sample screen from 2x2 and 3x3 grids. In the 2x2 grid case (Figure 4(a)) four cameras record the content of the environment while the video recordings from nine cameras render on the 3x3 grid (Figure 4(b)). The red rectangular boxes that appear in some of the videos represent the secure or forbidden areas whose violation causes



(a) (b)

Figure 4: Sample videos from the simulated factory-like environment representing (a) 2x2 grid and (b) 3x3 grid scenarios. The red rectangular boxes in the screens represent dangerous or forbidden areas. An event occurs when a moving object attempts to enter these areas.

a security event. Notice that since there are more than one forbidden area in 2x2 and 3x3 cases there may be multiple event occurrences of events. When two or three events occur almost simultaneously a double-event or a triple-event is reported, respectively.

Each file containing the 2x2 or 3x3 video screens has an associated log. This log includes vital information about the contents of the video and the number of possible security violations as well as their exact timing. This database is deployed in a computer lab on multiple client workstations.

After deploying the software packages in the computer lab, each workstation runs the evaluator software. This software is an advanced Graphical User Interface (GUI) which shows the videos to the users in a random manner. Each user watches the video and detects possible security issues contained in one or more of the videos within the 2x2 or 3x3 grids. The user may report the violation of secure regions that he/she detected by interacting with the Evaluator program. The Evaluator software records the reactions of its users in a database associating each user with his/her recorded data files.

Similar videos were created with the inclusion of the intent recognition module. The purpose of this module is to observe the videos and to apply a computational process to detect the intent of each animated object. Once the module gathers enough evidence about the possibility of a security issue within a video in the 2x2 or 3x3 screen grids, it will highlight the appropriate grid location. Therefore, it directs the intention of the user to the location containing the possible issue.

Figure 5 shows two examples of a 3x3 grid with the intent recognition aid embedded to highlight screens with a high likelihood of an event occurrence. In figure 5(a) a single event is about to occur with a high probability in the next few seconds in the video shown in the middle screen. A triple event (three almost simultaneous events) occur in the top-center, bottom-left and bottom-right screens in figure 5(b).

By employing these software packages user performances in detecting single, double, and triple security breaches can be recorded, evaluated and compared for each scenario.

Since the aid of the intent recognition redirects the operators' attention to the highlighted screen, simultaneous observation of multiple screens transforms into watching one video screen (out of 4 or 9) at a time. This is a big step toward achieving system independence performance - regardless of the number of screens observed.

**Research Design and Statistical Analysis**

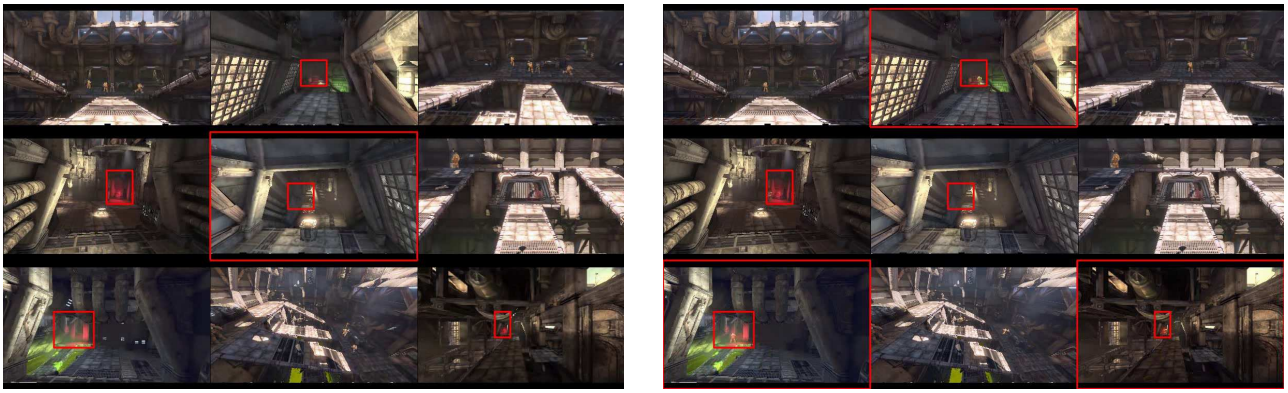
The study's design was a mixed-design experiment. There were two phases to the experiment. In phase 1 there was no intent recognition aid. In phase 2 there was an intent recognition aid, the red perimeter of a square. In each phase each participant was presented with single events, double events and triple events first with four videos (2 x 2) and then with nine videos (3 x 3).

There were five dependent variables:

1. SDR: percentage of single events detected
2. DDR: percentage of double events detected
3. TDR: percentage of triple events detected
4. FP: total number of false positives (participant detected a non-existent event.)
5. FN: total number of false negatives (participant missed a significant event.)

The within-subjects factor was time (phase 1 versus phase 2). The between-subjects factor was the number of videos (4 videos versus 9 videos).

A mixed-design multivariate analysis of variance (MANOVA) was performed to determine if there was a statistically significant difference by time (phase 1 versus phase 2) or number of videos (4 versus 9) on the dependent variables or if there was a statistically significant interaction between time and the number of videos on the dependent variables.



(a) (b)

Figure 5: Highlighted sample videos from the second phase of the simulated factory-like environment trials representing the recognition of (a) a single event and (b) triple events. The red rectangular highlight boxes around each screen represent the video with a high likelihood of an event occurrence.


Figure 6: A typical example of a participant's response in detection of single(green lines), double (black lines), and triple events (red lines) in phase one, without the intent recognition aid.

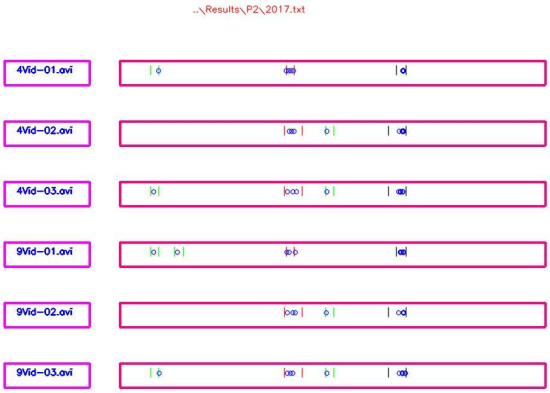



Figure 7: A typical example of a participant's response in detection of single(green lines), double (black lines), and triple events (red lines) in phase two, with intent recognition aid.

## 4. EXPERIMENTAL RESULTS

Figure 6 shows a sample response from a participant in phase one of the experiment. In the figure, each blue circle represents the detection of an event by the participant. The intervals shown between black, blue, and green lines show the approximate timing of single, double, and triple events, respectively. The participant ID number is shown on top of the chart.

According to our naming convention, the number at the beginning of the video name is the number of cameras used to make the surveillance video - i.e 4 cameras for 2x2 grid and 9 for 3x3 grids. The number at the end of the video name indicates the scenario number.

By observing the participant's response in phase one (figure 6), it is observed that the single events at the beginning of 4Vid-01, 9Vid-01, and 9Vid-03 are missed by the participant. At least one of the events within the triple event examples are also missed by the participant. According to figure 6, there are 4 false negatives (blue circles outside of the event intervals) in 2x2 camera scenarios and 3 in 3x3 camera scenarios. Also notice the false positives (extra circles within the event intervals) events occurring at the end of the timeline of each example.

In figure 7 the response from the same participant in phase two of the trial is shown. In the second phase the intent recognition aid processed the video from each camera and detected the potential events within each video separately. Once a potential event was detected by the intent recognition aid, its respective location was highlighted to direct the participant's attention to area in which the event would potentially occur.

In comparing the responses from figures 6 and 7, it can be observed that the detection of events are more accurate in phase two compared to phase one. From figure 7, the participant's response did not have any false positive (extra circles within the event intervals) or false negatives (blue circles outside of the event intervals). Also with the exception of the triple event in the 9Vid-01 scenario the partic-

Table 1: Follow-up ANOVAs By Phase

| Group | F(1,106) | p | Partial $\eta^2$ |
|---|---|---|---|
| SDR | 34.20 | < 0.001 | 0.24 |
| DDR | 60.11 | < 0.001 | 0.36 |
| TDR | 128.68 | < 0.001 | 0.55 |
| FP | 54.87 | < 0.001 | 0.34 |
| FN | 13.624 | < 0.001 | 0.56 |

ipant detected all other events with the help of the intent recognition aid in phase two of the study.

The mixed-design MANOVA revealed a statistically significant difference between time (phase 1 versus phase 2), $F(5, 102) = 59.89$, $p < .001$, partial $\eta^2 = .75$. Also there was a statistically significant difference between number of videos (4 versus 9), $F(5, 102) = 5.22$, $p < .001$, partial $\eta^2 = .20$ as well as a statistically significant interaction between time and number of videos, $F(5, 102) = 4.85$, $p = .001$, partial $\eta^2 = .19$. This interaction effect indicated that the difference between the number of videos on the linear combination of the five dependent variables was different at phase 1 than it was at phase 2.

Follow-up ANOVAs revealed that the significant change from phase 1 to phase 2 was statistically significant for each of the five dependent variables (see Table 1).

Examination of the means (see Table 3) and profile plots of the means for each dependent variable indicate that the lines converge or cross for all five dependent variables. However, follow-up univariate $F$ statistics indicated that the only statistically significant interactions between time and number of videos were for the following dependent variables ( see Table 2):

- TDR (percentage of triple events detected by the participant) with: $F(1, 106) = 10.43$, $p = .002$, and partial $\eta^2 = .09$
- FN (total number of false negatives) with: $F(1, 106) = 16.78$, $p < .001$, and partial $\eta^2 = .14$

Examination of the means indicated that for the dependent variable TDR the mean for four videos was higher than the mean for nine videos in phase 1 but not in phase 2 where the mean for both sets of video was equal but higher. Examination of the means indicated that for dependent variable FN the mean for 9 videos was much higher than the mean for four videos in phase 1 but not in phase 2 where the means for both sets of video were much lower and almost equal.

The statistically significant increase and convergence of the triple detection rate (TDR) means indicated the powerful effect of intent recognition on triple events detected for both 2x2 and 3x3 videos. The statistically significant decrease and convergence of false negatives (FN) means indicated the powerful effect of intent recognition on decreasing the number of false negatives for both cases.

Table 2: Significant Interactions Between Phase and Number of Videos

| Group | F(1, 106) | p | Partial $\eta^2$ |
|---|---|---|---|
| TDR | 10.43 | 0.002 | 0.09 |
| FN | 16.78 | < 0.001 | 0.14 |

Table 3: Descriptive Statistics

| Group | | Mean | $\sigma$ | N |
|---|---|---|---|---|
| SDR1 | 4 Videos | 0.65 | 0.26 | 54 |
| | 9 Videos | 0.62 | 0.21 | 54 |
| | Total | 0.64 | 0.23 | 108 |
| SDR2 | 4 Videos | 0.79 | 0.22 | 54 |
| | 9 Videos | 0.80 | 0.21 | 54 |
| | Total | 0.79 | 0.21 | 108 |
| DDR1 | 4 Videos | 0.83 | 0.22 | 54 |
| | 9 Videos | 0.78 | 0.25 | 54 |
| | Total | 0.81 | 0.24 | 108 |
| DDR2 | 4 Videos | 0.99 | 0.04 | 54 |
| | 9 Videos | 0.98 | 0.11 | 54 |
| | Total | 0.98 | 0.08 | 108 |
| TDR1 | 4 Videos | 0.55 | 0.34 | 54 |
| | 9 Videos | 0.33 | 0.34 | 54 |
| | Total | 0.44 | 0.35 | 108 |
| TDR2 | 4 Videos | 0.83 | 0.27 | 54 |
| | 9 Videos | 0.83 | 0.30 | 54 |
| | Total | 0.83 | 0.29 | 108 |
| FP1 | 4 Videos | 9.31 | 6.00 | 54 |
| | 9 Videos | 8.96 | 9.09 | 54 |
| | Total | 9.14 | 7.67 | 108 |
| FP2 | 4 Videos | 4.96 | 3.94 | 54 |
| | 9 Videos | 5.57 | 5.50 | 54 |
| | Total | 5.27 | 4.77 | 108 |
| FN1 | 4 Videos | 3.5 | 2.33 | 54 |
| | 9 Videos | 5.78 | 2.82 | 54 |
| | Total | 4.64 | 2.82 | 108 |
| FN2 | 4 Videos | 1.67 | 2.28 | 54 |
| | 9 Videos | 1.96 | 2.36 | 54 |
| | Total | 1.81 | 2.32 | 108 |

## 5. CONCLUSIONS AND FUTURE WORK

As shown in the experimental results section and Table 3, the intent recognition improved the event detection rate by individuals performing video surveillance tasks in all experiments. The event detection rates in different surveillance settings, i.e. 2x2 and 3x3 screen systems are comparable with the integration of the intent recognition into the system. This reinforces our hypothesis that the intent recognition reduces the cognitive burden of watching multiple screens and results in a system independent surveillance application.

In this paper we quantified the performance of humans in real-time video surveillance applications. We designed

a simulated factory-like environment in an advanced 3D graphical rendering software package. The video surveillance tasks were done by 54 undergraduate participants who watched 2x2 and 3x3 videos taken from 4 and 9 simulated cameras respectively.

In phase one of the study the participants observed the surveillance videos to detect single, double and triple events. An intent recognition framework was developed and implemented in the second phase of the study to highlight video screens with a high likelihood of an event occurrence. We tested our hypothesis that the intent recognition will help improve human video surveillance performance. Moreover, we quantified the statistical significance of the intent recognition aid in improving surveillance performance in this study.

With the effects of the intent recognition mechanism on improving human video surveillance tested in this paper, potential future directions for this study expand over a number of possible areas. Assistive technologies are a prime field for the expansion of this work. The intent recognition may be applied to the human computer interaction and graphical user interfaces to help individuals with physical disabilities to interact with computers in a more efficient manner.

Moreover, this technique can be applied to detect emergent patterns from vast sets of data. This may lead the techniques to detect potential disease causing genes and human-pathogen inter-genomic interactions before such interactions occur in a subgroup of human population.

## 6. ACKNOWLEDGEMENTS

The implementation of simulated environments used in this study was done in Unreal Development Kit (UDK) by Epic Games. The authors would also like to thank the participants who performed video surveillance tasks designed in the two phases of this study.

## REFERENCES

[1] K. Katsur, N. Sulman, T. Sanocki, D. Goldgof, How effective is human video srveillance perfromance?, in: Proceedings of the International Conference on Pattern Recognition, Tampa, FL, 2008.

[2] A. Tavakkoil, R. Kelley, C. King, M. Nicolescu, M. Nicolescu, G. Bebis, A visual tracking framework for intent recognition in videos., in: Proceedings of the International Conference on Pattern Recognition, Tampa, FL, 2008.

[3] R. Kelley, C. King, A. Tavakkoli, M. Nicolescu, M. Nicolescu, G. Bebis, An architecture for understanding intent using a novel hidden markov models formulation, International Journal of Humanoid Robotics 5 (22).

[4] R. Kelley, A. Tavakkoli, C. King, M. Nicolescu, M. Nicolescu, Understanding activities and intentions for human-robot interaction, In-Tech.

[5] A. Tavakkoil, M. Nicolescu, G. Bebis, A spatio-spectral algorithm for robust and scalable object tracking in videos, in: Proceedings of the 6th International Symposium on Visual Computing., Las Vegas, NV, 2010.

[6] A. Elgammal, R. Duraiswami, L. Harwood, D. nd Davis, Background and foreground modeling using nonparametric kernel density estimation for visual surveillance., in: Proceedings of the IEEE., 2002, pp. 1151–1163.

[7] A. Tavakkoil, R. Kelley, C. King, M. Nicolescu, M. Nicolescu, G. Bebis, A vision based architecutre for intent recognition., in: Proceedings of the 4th International Symposium on Visual Computing., Las Vegas, NV, 2008.

[8] C. Stauffer, W. Grimson, Learning patterns of activity using real-time tracking, IEEE Transaction on Pattern Recognition and Machine Intelligence 22.

[9] L. Li, W. Huan, I. Gu, Q. Tian, Statistical modeling of complex background for foreground object detection, IEEE Transaction on Image Processing 23.

[10] A. Tavakkoil, M. Nicolescu, M. Nicolescu, G. Bebis, Incremental svdd training: Improving efficiency of background modeling in videos, in: Proceedings of the 10th IASTED International Conference on Signal and Image Processing., Kona, HI, 2008.

[11] Y. ar Shalom, Tracking and data association, Academic Press Professional, Inc.

[12] G. Kitagawa, Non-gaussian state-space modeling of nonstationary time series, Journal of American Statistical Association 82.

[13] M. Isard, A. Blake, Condensation– conditional density propagation for visual tracking, International Journal of Computer Vision. 29 (1) (1998) 5–28.

[14] L. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, Proceedings of the IEEE 77.

[15] D. Comaniciu, V. Ramesh, P. Meer, Kernel based object tracking, IEEE Transaction on Pattern Recognition and Machine Intelligence 25.

[16] K. Sankaranarayanan, J. W. Davis, Attention-based target localization using multiple instance learning., in: Proceedings of the 6th International Symposium on Visual Computing, Las Vegas, NV, 2010.

[17] K. Paputsakis, A. Argyros, Object tracking in a closed loop., in: Proceedings of the 6th International Symposium on Visual Computing, Las Vegas, NV, 2010.

[18] D. Permack, G. Woodruff, Does the Chimpanzee have a Theory of Mind?, Behavioral and Brain Sciences 1 (4) (1978) 515–526.

[19] A. Gopnick, A. Moore, Changing your views: how understanding visual perception can lead to a new theory of mind, in: C. Lewis, P. Mitchell (Eds.), Children's Early Understanding of Mind, Lawrence Erlbaum Press, 1994, pp. 157–181.

[20] D. Baldwin, J. Baird, Discerning intentions in dynamic human action, Trends in Cognitive Sciences 5 (4) (2001) 171–178.

[21] J. Gray, C. Breazeal, M. Berlin, A. Brooks, J. Liberman, Action parcing and goal inference using self as simulator, in: Proceedings of the 14th IEEE International Workshop on Robot and Human Interactive Communication.