Recognition of Abnormal Human Behavior using Kinect Case Study: Tehran Metro Station

Farzam Ghanbarnezhad*, Salman Hosseini2 and Motahareh Hosseini3

1Computer Science, Artificial Intelligence, Semnan University, Semnan, Iran: ghanbarnezhad@gmail.com
2Information Technology, Semnan University, Semnan, Iran; salman.ho.98@yahoo.com
3Business Administration - Computer Information Systems, Colorado State University, Fort Collins, USA: mo.hosseini.gh@gmail.com

Abstract

Objectives: The ability of detecting people and understanding their behaviors for recognition or detection of special event has attracted significant research interest in recent years. Methods: The main purpose of this study focuses on abnormal human behavior detection by using Microsoft's Kinect camera. By utilizing Kinect, the system tracks all persons in Metro station and calculates their position relations, which provide a more effective way to analyze human's behavior. Results: It first applies Kinect technology of Microsoft in the field of intelligent monitoring. By taking use of the depth information, the system recognizes unusual behavior and by face detection and analysis, the system of face-hiding behavior. The time people spend in Metro station is recorded and spending too much time is also regarded as an abnormal behavior. Applications: Experiments show that the system is very robust for abnormal behavior recognition and provides a new way for preventing Metro station crimes.

Keywords: Abnormal Human, Behavior, Kinect, Recognition

1. Introduction

Understanding human behaviors is a challenging problem in computer vision and automated video-based surveillance systems for real-time human behavior analysis provides a reliable solution of detecting the abnormal behaviors and events among our environment. The ability of detecting people automatically and understanding their behaviors is a key of intelligent video surveillance systems and the technical challenges issues are needed to reliably detect and track moving targets within a highly dynamic background and interpreting targets' behavioral patterns. The increase of population in societies and responsibility of governments to create a social security causes the demand for such systems increases day by day. A system that can monitor the unusual behavior of humans, especially in crowded places can reduce crime and increase security in society. It detects human's abnormal behaviors and considered people who try to prevent their faces from being monitored and people who spend too much time at Metro station. This experiment, unlike previous methods that used ordinary camcorder, uses Microsoft's Kinect camera. By utilizing Kinect, the system tracks all persons in Metro station and calculates their position relations, which provides a more effective way to analyze human's behavior. When the system detects the above mentioned two abnormal behaviors, it gives security staffs of the Metro station center an alarm. It also records these abnormal behaviors separately to help police find suspects.

Experiments are performed in a large room with about 30 people to resemble Metro station. The system consists of three abnormal behaviors recognition. The schematic diagram by Kinect can monitor and save positions behavior, face detection and people when spend too much wandering in simulated environment. The system classifies the two behaviors and records their occurrence time in report.
forms. It will be very convenient and high-efficient for police to look for suspects from the videos. As mentioned one of the most important research area which Kinect widely used is in recognizing human behavior. There is an increasing demand for smarter video surveillance of public and private space using intelligent vision systems which can distinguish what is semantically meaningful to the human observer as ‘normal’ and ‘abnormal’ behaviors. Regarding human recognition behavior there have been many research works, most of the behavior recognition are detecting the normal behavior of human in different situation and using statistic cameras. In this study human abnormal behavior has been achieved by using Kinect. The abnormal behavior of human at Metro station is considered for this study. Metro station is a place there can bring some crimes.

1.1 Kinect Technology
Microsoft Kinect, a low-cost motion sensing device, enables users to interact with computers or game consoles (XBOX) naturally through gestures and spoken. With the introduction of the free Microsoft Kinect SDK in 2011, the Kinect technology opened a huge door for developing other applications beyond Xbox games. Major Kinect application categories span from healthcare, to education, retail, training, gaming, robotics control, natural user interface, sign language recognition as well as 3D reconstruction, which is a great technology for the 3D printing revolution. Except for 3D reconstruction applications, virtually all other Kinect applications require motion recognition so that the semantics of a human gesture or action can be interpreted automatically. Hence, motion recognition is the fundamental enabling technology. The innovative technology behind Kinect is a combination of hardware and software contained within the Kinect sensor accessory that can be added to any existing XBOX 360. To use the features of the hardware, it must also use the proprietary layer of Kinect software that enables body and voice recognition from the Kinect sensor.

1.2 Kinect Specification
According to Microsoft website, a Kinect for Windows sensor contains:
- An RGB camera that stores three channel data in a 1280 x 960 resolution. This makes capturing a color image possible.
- An Infrared (IR) emitter and an IR depth sensor. The emitter emits Infrared light beams and the depth sensor reads the IR beams reflected back to the sensor. The reflected beams are converted into depth information measuring the distance between an object and the sensor. This makes capturing a depth image possible.
- A multi-array microphone, which contains four microphones for capturing sound. Because there are four microphones, it is possible to record audio as well as find the location of the sound source and the direction of the audio wave.
- A 3-axis accelerometer configured for a 2G range, where G is the acceleration due to gravity. It is possible to use the accelerometer to determine the current orientation of the Kinect Figure 1.

2. Historical Development of Different Sensors for Monitoring
Researchers have used wearable sensors in an attempt to obtain accurate data from people being monitored. An example of this type of sensor involves an RFID chip embedded into a wearable bracelet. This system involves adding RFID tags to many common household objects or areas and works by recording any time the bracelet comes into contact with these objects. This means the sensor can detect when a person comes into contact with objects such as a light switch or broom, from which activity recognition can be determined. These types of wearable sensors have provided researchers with promising results, however the problem with using these attachable sensors is that people need to wear them, therefore, in secretly detecting behaviors is not applicable. To deal with this problem and to minimize the level of human interaction needed to
monitor, a sensor that monitors the electrical impulses of a residential power line has been implemented by some researchers. In this system, the sensor detects electrical impulses when an electrical appliance is turned on or off. These impulses are unique to each electrical appliance which allows researchers to track when the various appliances or lights are in use. This data is then used to determine the actions of a person. Drawbacks of this are that only the usage of electrical appliances and fixtures can be monitored. There are many activities where the only electrical component is to turn on a light in a particular room and if enough natural light is present that might also not occur. A common set of sensors that are being used in the field of activity recognition and abnormality detection are standard cameras. Cameras are cheap and easily accessible and when used to monitor a person over a long period of time they do not impose a burden on the subject that is seen with the wearable sensors that were previously discussed. Using a standard RGB video requires the use of processing techniques, such as background subtraction, to separate the people in the foreground of the image from the background of the image. Because of drawbacks of RFID this research will be using the Kinect as a means of data acquisition; this means the data received from this device may lead to more accurate results in a range of conditions. As well as the ability to obtain standard RGB images like that of a video camera, it is the Kinect’s Infrared sensor that separates it from many other types of sensors. The Infrared sensor allows the Kinect to obtain the distance between itself and any object within its range. Unlike embedded sensors, the Kinect can be easily added to an existing environment without installation cost.

3. Recognition Techniques and Methods of Human Behaviors

Different types of techniques and methods have been developed to assist automated activity recognition and in turn abnormality detection. Their techniques have two classifications: Single layered approaches and Hierarchical approaches. A Single-layered approach is one that recognizes a sequence of actions directly from a video and matches them to a suitable class. A Hierarchical approach differs in that it uses the recognition of several smaller actions to recognize a larger or more complex activity and represent it at a higher level. The single-layered approach is categorized into space-time approaches and sequential approaches.

3.1 Space-Time Approaches

Space-time approaches are those which model video data as a 3-D volume, using time as the added dimension. This is done using the 2-D frames from a video and linking them along a time axis in order to obtain the 3-D volume. Other forms of space-time approaches include space-time trajectories and space-time features. Trajectory-based approaches are those which track the path of a person’s joints, e.g. which means as a person moves throughout a video, the position of their joints are recorded into a 3-D representation. Feature-based approaches deal with 3-D volume, like that obtained in the space time volume approach, but involve the extraction of specific features from this volume which can be used to identify certain activities. In1 employed a trajectory based approach which used template matching for activity recognition. Their work was based on tracking the position of hand movements, which was used to create the trajectories. Subsequently, utilized a space-time feature approach in which statistical modeling was used to detect a series of human actions, facial expressions and hand gestures.

3.2 Sequential Approaches

Sequential approaches differ to space-time approaches in that they represent human activity as a sequence of features and recognize these activities by searching for this sequence. Sequential approaches are further categorized into exemplar-based and state model-based.

3.3 Hierarchical Approaches

The hierarchical approach is further categorized into statistical, syntactic and description based approaches.

3.3.1 Statistical Approaches

Statistical approaches use multi-layered graphs to categories sequential activities. The first layer is usually similar to the single-layered sequential approach in that actions are recognized from a sequence of features. An example of a hierarchical statistical approach can be seen in the work of in2, which involves a system for human actions that uses Hidden Markov Model to detect activities. In2 another example of a hierarchical statistical approach,
however this work focuses on group interactions and uses a Bayesian model to recognize activities.

### 3.3.2 Syntactic Approaches

Syntactic approaches are those which model human activity as a string of actions. Similar to the statistical approach, the actions are the small tasks which make up an activity and must be recognized first. An example of a hierarchical syntactic approach can be seen in the work of 3, who developed a system using attribute grammars to detect abnormalities in a business parking lot.

### 4. Description Based

In description based approaches the structure of activities is characterized and sub activities are detected based on a predefined activity relationship model. A system created by in\(^1\) demonstrate a description based approach can be used to recognize group interactions as well as interactions between multiple groups, eg. two groups fighting.

### 5. Behavior Analysis with the Kinect

The recognition of behavior is used in different research areas, such as surveillance. In our application we are monitoring a single person to detect whether he is showing abnormal. The Kinect is a motion-capture device that provides the developer with the 3D location and skeleton posture of the user. The Kinect sensor is also equipped with an RGB camera and microphone array. Most research that uses the Kinect as sensor uses the skeleton representation as input for the recognition of activities. In recognize everyday activities such as “brushing teeth and working on computer” from recorded skeleton data, even when the person was not seen before. In use the skeleton information indirectly, by computing the 3D bounding box around a detected person to recognize a fall. In estimate the respiratory rate based on the expansion and contraction of the chest area of the person tracking across multiple cameras. Even though the Kinect is equipped with good tracking functionality, at the moment of writing we cannot yet track a person across multiple cameras. The research by in\(^4\) monitors the behavior of children across multiple sensors. While in try to detect anger and bullying behavior in children, in\(^5\) try to monitor children standout that might be of risk of mental illnesses as autism and obsessive compulsory disorder as they display subtle behavior divergences. Both projects are interesting for our application, but are still in an exploratory state. In this section I presented some research projects that use the Kinect as sensor. Most of them use the skeleton data to get the location and body position of a person that is tracked\(^6\).

### 6. Methodology

The Microsoft Kinect was released in November 2010 for the XBOX 360 gaming console. Soon after it was released, developers were able to connect it to a computer and create their own applications with the Kinect. Range of the Kinect horizontally of view of the Kinect is 57 and the vertically of view is 43. The Kinect can also be tilted in the vertical axis, either a maximum of 27 up or down. Therefore we can place the Kinect at the door or below ceiling as well. The Kinect has an optimal range between 1.2 and 3.5 meters, though it can also track people between 0.8 and 6 meters. Tests in the simulated room showed that this range was sufficient to capture a view of the entire room with two Kinect sensors. When the distance to the sensor increases, the depth measurements get less accurate.

### 7. Simulated Environment

The system consists of three abnormal behaviors recognition. By using Kinect, it detects and tracks all persons in the specified visibility range in Metro station. It can also get all persons position information from Kinect. By utilizing the position information, it recognizes peeping behavior. By face detection, it recognizes face-hiding behavior. Besides, when people spend too much time wandering in a Metro station, it also can be recognized as an abnormal behavior. The system classifies the three behaviors and records their occurrence time in report forms. It will be very convenient and high-efficient for police to look for suspects from the videos.

### 8. Implementation

Kinect is a motion sensing input device by Microsoft for the XBOX 360 video game console and Windows PCs. By utilizing Kinect and Kinect SDK, the system can easily track people in the specified environment of Metro sta-
tion and obtain their depth information. In this demo, the size of the depth map is 320×240 pixels. Each pixel represents the Cartesian distance, in millimeters, from the camera plane to the nearest object at that particular X and Y coordinate, as shown in a pixel value of 0 indicates that the sensor did not fit any objects within its range at that location. Their distance in Z direction can be calculated by subtracting their depth value. The x and y coordinates on image can be turned into the real X and Y coordinates in Kinect coordinate space by Kinect SDK. Then the distance of two human is computed by the following equation:

\[ \text{Dis} = \sqrt{(z_1 - z_2)^2 + (x_1 - x_2)^2} \]

9. Face-Hiding Recognition

By using the RGB-D camera of Kinect, the system can easily track user’s head and extract his head area. Then, the system detects face in the head area. For face detection, the Face SDK of Lux and was adopted. The Face SDK has robust frontal face detection and allows -30...30 degrees of in-plane head rotation, -10...10 degrees out of plane rotation. Face occlusion, especially wearing sunglass or mask can also be recognized perfectly. In addition, the time customers spend in Metro station can be recorded correctly. All abnormal behaviors above will be alarmed and written into report respectively.

10. Other Abnormal Behaviors

Also there are some other abnormal behaviors, such as wandering, destroying equipment in Metro station or fighting others. These behaviors usually cost criminals more time. Thus, when a customer spends too much time in Metro station, the system will give an alarm and remind workers in monitoring centers.

11. Starting the Recording

During the recording we display the depth data from the Kinect sensor, together with the skeleton visualization and the distance to the sensor. The depth-data is displayed according to the color scheme. We only store the skeleton data of a tracked person. For our recognition approach we only require the location of the person.

12. Data Storing

In order to reuse the data we store it, both as a serialized object and in a .csv (comma-separated values) file. Serialization transforms an object into a format that can be stored. It can also be deserialized, allowing us to access the object as before. The .csv file can be read by other programs as well and stores the data in a human readable format. We store general information about the size of the room and the angle of the recording Kinect. For each frame we store a time stamp, the location of the skeleton and the location and tracking state of each joint.

13. Demonstration

We have conducted several experiments to demonstrate the system. The result shows that the system can detect abnormal behavior in real-time and give alarms. The system can detect front faces robustly and allow -30...30 degrees of in-plane head rotation, -10...10 degrees out of plane rotation. Face occlusion, especially wearing sunglass or mask can also be recognized perfectly. In addition, the time customers spend in Metro station can be recorded correctly. All abnormal behaviors above will be alarmed and written into report respectively.

14. Conclusion

This demo has presented a real-time system for monitoring abnormal behaviors in Metro station. It first applies Kinect technology of Microsoft in the field of intelligent monitoring. By taking use of the depth information, the system recognizes unusual behavior. By face detection and analysis, the system finds face-hiding behavior. The time people spend in Metro station is recorded. Spending too much time is also regarded as an abnormal behavior. Experiments show that the system is very robust for abnormal behavior recognition. It provides a new way for preventing Metro station crimes.

15. References

1. Nguyen NT, Phung DQ, Venkatesh S, Bui H. Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. 2005 Jul; 2:955–960.
2. Dai P, Di H, Dong L, Tao L, Xu G. Group interaction analysis in dynamic context. IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics). 2009 Feb; 39(1):34–42.
3. Recognition of high-level group activities based on activities of individual members. 2008. Available from: http://ieeexplore.ieee.org/document/4544065/
4. Yao A, Gall J, Van Gool L. Coupled action recognition and pose estimation from multiple views. International Journal of Computer Vision. 2012 Oct; 100(1):16–37.

5. Martin R, Sujatha S. Bottom-up approach of modeling human decision making for building intelligent agents. Indian Journal of Science and Technology. 2016 Jan; 9(4):1–5.