Recognizing Unusual Behaviour for Intelligent Space

Andor Gaudia, Péter Szemes, Béla Takarics, Péter Korondi

Budapest University of Technology and Economics
gaudia@get.bme.hu

Abstract: In these days when we go to public places like banks or post-offices, we always have to reckon with the risk of armed robbery, or terrorist attack. These places are protected by special security crew, but the increasing demand on improved security requires redundant and failsafe methods which eliminates the human factor. In this paper an Intelligent Space based project is introduced which can detect unusual human behaviour in the monitored environment using computer vision.

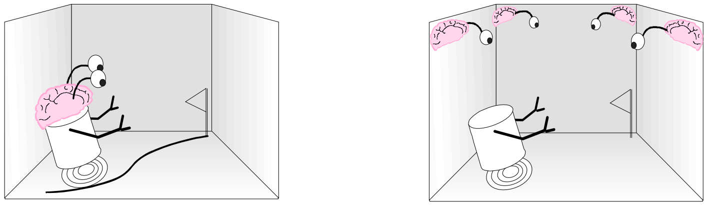
Keywords: Intelligent Space, image processing, genetic algorithm, background removal, skin color, genetic-search

1 Introduction

1.1 A Brief History of the Intelligent Space

Hashimoto Lab. in The University of Tokyo has proposed 'Intelligent Space' since 1996. At the beginning it consisted of two sets of vision cameras and computers with a home made 3D tracking software, this was written in C and tcl/tk under Linux. Later, a large-sized video projector (100 inches) was added to the Intelligent Space as an actuator. Mobile robots were located in the Intelligent Space for supporting people as well as for being supported. Vision cameras and computers sets were arranged around an entire room and it changed into the Intelligent Space.

Conventionally, there is a trend to increase the intelligence of a robot (agent) operating in a limited area. The Intelligent Space concept is the opposite of this trend. The surrounding space has sensors and intelligence instead of the robot (agent). A robot without any sensor or own intelligence can operate in an Intelligent Space. The difference of the conventional and Intelligent Space concept is shown in Figure 1.



Conventional concept

Intelligent Space concept

Figure 1
Difference between the conventional and Intelligent Space concepts

1.2 What is Intelligent Space?

Often, science fiction movies become good references for the actual engineering. Some unbelievable systems in film have appeared with progress of technology such as space rockets, robots. In our laboratory, a system, which may be seen in the films, has been realizing. It is called 'Intelligent Space Project'. Before explaining what it exactly is, I would like to introduce a few movie stories. In the movie '2001: A Space Odyssey', a computer named 'HAL' has high intelligence. The HAL can watch human's activity with its distributed cameras and control subordinate systems as expanded actuators of it. Another SF movie, titled 'Demon Seed', a brilliant scientist of the future creates a computer named 'Proteus' with almost limitless intelligence. However, the Proteus tried to produce offspring and it hinders all the people who plan to get rid of the Proteus. As HAL did, the Proteus utilizes all the electrical systems in the house as its parts. Unfortunately both of the movies are telling us the fear of technology when the machine gets high intelligence. However, we should notice the intelligent systems in those movies. Some people may say these are ubiquitous computing, but we recognized those systems as intelligent environment. Such intelligent environments are able to watch what is happening in them, build a model of them, communicate with their inhabitants and act based on decisions they make. Especially the capability of the environment to act as a context-sensitive user interface (e.g. to respond to gestures) and react in certain situations (e.g. accidents, intruders) promises a range of application scenarios such as intelligent hospital rooms, office, factory, asylum for the aged, etc.

1.3 Concept of the Intelligent Space

Intelligence Space is a space (room, corridor or street), which has distributed sensory intelligence (various sensors, such as cameras and microphones with intelligence, haptic devices to manipulate the space) and it is equipped with actuators. Actuators are mainly used to provide information and physical support to the inhabitants. This is done by speakers, screens, pointing devices, switches or robots and slave devices inside the space. The various devices of sensory intelligence cooperate with each other autonomously, and the whole space has high intelligence. Each agent has sensory intelligence. The intelligent agent has to operate even if the outside environment changes, so it needs to switch its role autonomously. The agent knows its role and can support man. Intelligent Space re-composes the whole space from each agent's sensing information, and returns intuitive and intelligible reactions to man. In this way, Intelligent Space is the space where man and agents can act mutually.

1.4 DIND (Distributed Intelligent Networked Device)

In order to realize the Intelligent Space, sensors are located, which recognize space. However, we cannot change normal environment without considering economical and laborious problems. Moreover, appearance should be considered prudentially. Thus, it should be restricted to the range, which does not bring big influence on the existing environment. Based on these, Distributed Intelligent Network Device (DIND) is proposed (Figure 2), which is composed of three basic elements. The elements are sensor, processor (computer) and communication device. DIND is a small device based on three functions that the dynamic environment, which contains people and robots, is watched by the sensor, information is processed to be known easily by the clients by the processor and the DIND communicates with other DINDs through networks.

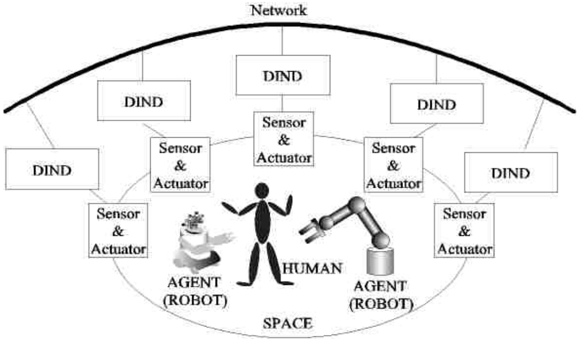


Figure 2
The structure of the DIND

2 General Description

Recognizing unusual behaviour in general is a rather complicated task and to complex to generalize it. In Hollywood movies a well known situation can be seen: the robber enters the bank, shouts and makes the customers to lie on the floor or asks them to keep their hands above their head on a visible place; This famous situation gave us the idea to focus on special body positions and try to recognize their meaning.

Different human behaviour forms were studied at different places and with a few exceptions a model was created based on our study. Individual persons and groups were studied and a probability table was created.

The most common standard behaviour forms at bank-offices were the followings:

- Standing, holding something in the hand
- Sitting on a chair
- Walking

The following table shows the different behaviour forms at a given time for 1 person, 2-3 people, and 90% of the people in the monitored area when they are doing their daily routine. In each cell the average activity duration is shown too.

	1 person	2-3 people	90% of the people
Standing	Often / > 1min	Often / > 40sec	Often / < 10sec
Walking	Often / 5 sec	Often / 5 sec	Rarely / -
Sitting	Often / > 5min	Often / > 3min	Sometimes / -
Crouching	Rarely / < 30 sec	Rarely / -	Never / -
Keeping their hands up (stretching)	Rarely / < 10 sec	Almost never / -	Never / -
Lying on the floor	Never / -	Never / -	Never / -

The following table shows the same probability values during the robbery.

	1 person	2-3 people	90% of the people
Standing	Often / > 1min	Often / > 1min	Often / > 1min
Walking	Often / 1-2sec	Rarely / -	Almost never / -
Sitting	Rarely / -	Rarely / -	Almost never / -
Crouching	Never / -	Never / -	Never / -
Keeping their hands up	Often / > 1min	Often / > 1min	Often / > 1min
Lying on the floor	Often / > 1min	Often / > 1min	Often / > 1min

According to the study, the positions of the hands were usually close to the body and the body was usually in vertical position. Special hand positions were noticed when somebody was stretching his limbs after getting tired but it wasn't significant.

Major differences can be found especially in those cases which are not common in public places like keeping the hands above the head for long time or lying on the floor. The problem of recognizing unusual behaviour can be simplified to recognizing lying people and standing people shapes with hands up.

The process of recognizing different body shapes can be separated into several sub-tasks, as follows:

- Creating a digital picture about the monitored environment
- Finding different objects on the picture
- Recognizing object shapes

Finding different objects on the picture

Camera systems are widely used in security systems, mainly for the purpose of recording the events to a video tape. These tapes can be examined after the robbery. Usually fixed cameras are used but sometimes pan/tilt camera systems can be found. The proposed system uses fixed color cameras with fixed zoom; the camera automatically adopts to the lighting conditions.

The process of finding separate objects on the picture needs the following steps to be performed:

- Separating objects from the background
- Separating objects from each other
- Interpreting the object seen on the picture

2.1 Background Substraction

The first and most trivial idea consists of making a background subtraction in order to keep only the user's body on the image.

The aim is to distinguish what we call the foreground and the background. In this chapter the term "background" stands for a set of motionless in age of pixels, that is, pixels that do not belong to any object moving in front of the camera. The most usual background subtraction techniques are:

- Average, median, running average,
- Mixture of Gaussians,
- Kernel density estimators,
- Mean shift (possibly optimized),

- SKDA (Sequential Kernel Density Approximation).

The first method was chosen because it is easy to implement and gives acceptable result at a reasonable speed.

2.1.1. The Chosen Method

The primary principle consists on comparing a known image, considered as a fully background to the current one. The result of the difference between the two is supposed to be the foreground. Before going further, we have to accept that the camera remains static during the whole process. This technique can be decomposed into three parts: initialization, extraction and update [1].

2.1.2. Initialization

First a static model of background has to be created, used as the reference image. The simplest background model assumes that, every background pixel brightness varies independently, according to normal distribution. The background characteristics can be calculated by accumulating several dozens of frames, as well as their squares, which means finding a sum of pixel values in the location S(x, y) and a sum of squares of the values Sq(x, y) for every pixel location. The mean is calculated as follows:

$$m(x,y) = \frac{S(x.y)}{N}, \tag{1}$$

where N is the number of frames collected. The standard deviation is calculated as:

$$\sigma(x,y) = \sqrt{\frac{S_q(x,y)}{N} \cdot \left(\frac{S(x,y)}{n}\right)^2} \; . \tag{2}$$

At the end of this calculation, we can obtain for each pixel its average value and its standard deviation.

2.1.3. Extraction

The brand new images are now compared to the static references previously created. Each pixel is compared one by one. The pixel, placed in (x, y) is regarded as belonging to a moving object if the following condition is met:

$$| m(x,y) - P(x,y) | > c \cdot \sigma(x,y) \tag{3}$$

where P is the current tested image and c is a constant, generally chosen between 2 and 4. Finally, we create a binary mask, called b(x, y) using this equation:

$$b(x,y)=\begin{cases}1 & if \quad |m(x,y)-P(x,y)| \geq c\cdot\sigma(x,y) \\ 0 & otherwise\end{cases}. \tag{4}$$

At this step, we easily understand why the camera does not suppose to move during the process and why should every object be put away from the camera for a few seconds, so that a whole image of the camera represents subsequent background observation.

Nevertheless, we have to think, this process is applied at the same time to each component of the colored pixel, that is let say red, green and blue if the color space is RGB, or hue, saturation and value, if it is HSV. To decide if a pixel belongs to the background we have to apply an OR function to these three signals. If 1 means the foreground and 0 the background, we can apply to each b(x, y) signal mask the following operation to obtain the final M(x, y) mask.

$$M(x,y)=b_r(x,y)\cup b_g(x,y)\cup b_b(x,y) \tag{5}$$

After this, any kind of image processing can be applied to mask in order to improve its quality: dilatation, erosion, convolution…Finally, we just have to apply this mask to the original input I(x, y) to obtain pure balck as background. The foreground F(x, y) is obtained in the following way:

$$F(x,y)=M(x,y)\cdot I(x,y) \tag{6}$$

2.1.4. Update

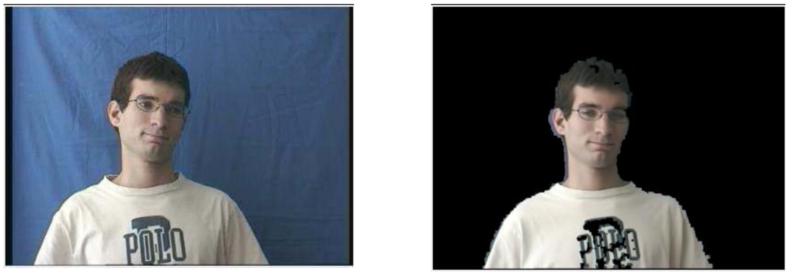
The above mentioned technique can be improved. First, it is reasonable to provide adaptation of background differencing model to changes in lightning conditions and background scenes, e.g. when some object is passing behind the front object.

The principle consists of making the static image change during time. This means the reference m(x, y) must be a mix of itself and the current background image.

$$m(x,y)=(1-\varepsilon)\cdot m(x,y)+\varepsilon\cdot I(x,y)\cdot \overline{M}(x,y) \tag{7}$$

$I(x,y)\cdot\overline{M}(x,y)$ represents the current background image (pixels belonging to the foreground are not updated) and ε represents the teaching rate of the image. This number must be low, between 0.25 and 0.5, if the static image is refreshed each 20 seconds for instance. This operation allows the system to support the sunrise or any change in the light conditions. Moreover, as the foreground is not considered, a quick change does not amend the static image.

Figure 3 illustrates the best result, we can reach actually with good lightning conditions.



(a) Original image

(b) Image after background removal

Figure 3
Background removal examles

2.2 Skin Color Detection

2.2.1 Basic Principles

The principle of the skin color recognition is very easy. According to researches, a pixel, which meats the following parameters is considered as a skin:

$$\begin{cases} \mathit{Hue} < 27 \\ 20 < \mathit{Saturation} < 80 \end{cases} \tag{8}$$

These conditions are very basic. Another model has been created thanks to many studies. The following improved conditions have been established [4]:

$$\begin{cases} S > 10 \\ V > 40 \\ S < -H - 0.1 \cdot V + 110 \\ S < 0.08 \cdot (100 - V) \cdot H + 0.5 \cdot V \\ 2 < H < 43 \end{cases} \tag{9}$$

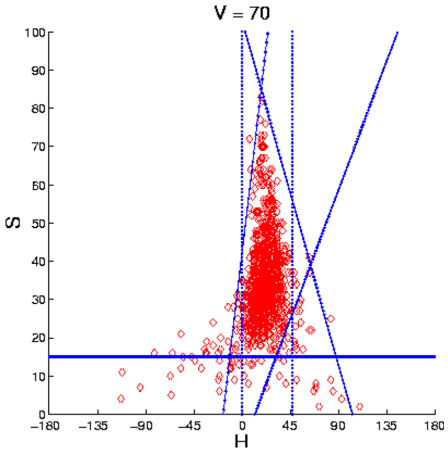
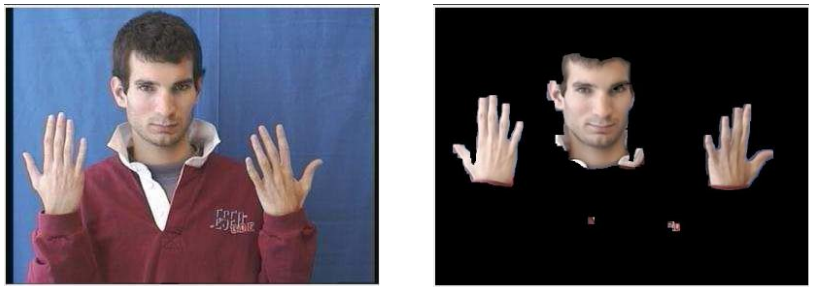


Figure 4
Pixel parameter conditions for skin color detection

Even this technique is basic; indeed, skin detection could be a whole paper subject itself. However, we shall see we can reach quite good level just using the second mentioned model. We must also take care to the values used to define the three variables. In the program, S and V are normalized to 1 and H is between 0 and 360, whereas sometimes S and V are between 0 and 100 and H is between -180 and +180.



(a) Original picture

(b) After skin recognition

Figure 5
Skin color detection examples

2.3 Matching the Objects

After separate objects are available they have to be identified; for a trained human eye and brain it's not a problem to identify humans and to find out their arm positions. At the current state of computing different ticks has to be done to achieve the aimed result. In computer vision usually a 3D model is used to generate a 2D picture; this process is called rendering. Our idea was to do it backwards and using a 2D image, generate the corresponding 3D model. 2D image does not contain enough information to perform this transformation directly, but estimations can be done, and the result could be good enough to be able to distinguish different body forms.

In computer vision different methods are available to implement 2D-3D conversion: cognitive approach, neural networks, genetic-search algorithms.

In this paper a genetic-search based approach will be used. Genetic-search algorithms are rather similar to other search algorithms, but the huge difference is that in this case the search tree is not available at the beginning. Search cases are generated on-the-fly using the calculated error of previous comparisons.

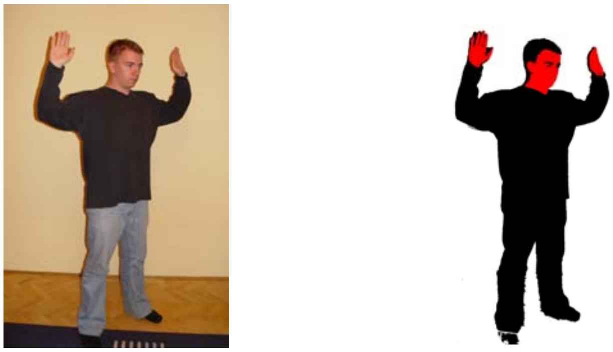
To implement this algorithm first of all a 3D model of the human body has to be created. The model not only contains the basic parts like arm, body, legs, head, but it contains additional information about the possible arm and leg positions. The model contains several constraints too: impossible body positions which violate the basic physical rules are not accepted.

The core of genetic algorithm is able to set up various leg and arm positions, and it is able to rotate the entire model in any possible way. If the generated model violates the rules then it is thrown away, but if it passes a 2D image is created using standard rendering. The rendered picture is compared to the reference object. The algorithm compares them and if the result meets the requirements the original 3D model is placed into the search tree. The algorithm will use this 3D model as a base to generate the new 3D model. If the model does not meet the requirements, it will be thrown away, it didn't survive the evolution.

The human body has infinite different arm/leg positions – the algorithm would need infinite time to choose the best 3D model – but the implementation will limit the acceptable arm/leg positions to a reasonable number.

The rendered image and the original object has to be compared. To be able to perform the comparison several steps has to be done to get accurate results:

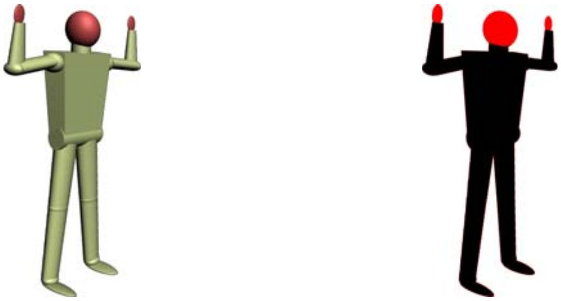
- scale the rendered image to match the size of the reference object
- convert both images into a B&W image
- position the rendered image to cover the reference image
- do a pixel-by-pixel comparison and count the match-ratio



(a) Original picture captured using a digital camera

(b) Processed image after background removal, B&W conversion and skin recognition

Figure 6
Captured image



(a) 3D model which survived the evolution

(b) Converted to B&W and skin color is marked with red

Figure 7
Generated 3D model

A better result can be achieved if in step 2 the skin color is converted into a special color – for example red – and on the 3D model the head and hands are colored the same way. In this case the genetic-search algorithm can calculate the center of gravity of the red blobs, measure the distance between them and provide a better heuristic function.

Conclusions

The proposed system integrates different techniques to achieve the final result. Background removal and skin detection algorithms were designed and implemented at Budapest University of Technology and Economics. The genetic algorithm is not yet implemented but the mathematical results and the experiments we did, show that it has a future and the system will be able to recognize unusual behaviour with high precision (according to calculations the hit ratio was >90%). The proposed system will be implemented as a DIND, so using several DINDs at different locations can improve the hit ratio up to 99.9%.

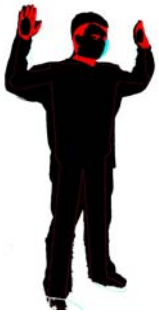


Figure 8

The generated and the captured placed on top of each other

Acknowledgement

The authors wish to thank the JSPS Fellowship program, National Science Research Fund (OTKA T046240), Control Research Group and János Bolyai Research Scholarship of Hungarian Academy of Science for their financial support and the support stemming from the Japanese-Hungarian Intergovernmental S & T Cooperation Programme.

References

[1] M. Piccardi. Background Subtraction Techniques: a Review. Faculty of Engineering UTS, 2004

[2] J.-H. Lee, H. Hashimoto. *Intelligent Space – Its Concepts and Contents*. Advanced Robotics Journal. Vol. 16, No. 4, 2002

[3] J. Ponce, D. A Forsyth. *Computer Vision: A Modern Approach.* Prentice Hall, 2002

[4] Xiu-Na Xu, Zhe-Ming Lu. *Real-time Face Detection Based on Skin Color Model.* Department of Automatic Test And Control, Harbin Institute of Technology, China

[5] G. Burdea and P. Coiffet. Virtual *Reality Technology.* New York: John Wiley & Sons, Inc, 1994