Multimedia Surveillance Station for Audio-Visual Objects Tracking with Mobile Robot

Alexander Bekierski, Emil Altimirski, Snejana Pleshkova
Department of Telecommunications
Technical University
Kliment Ohridski, 8 Sofia
aabbv@tu-sofia.bg, altimir@tu-sofia.bg, snegpl@tu-sofia.bg

Abstract: - The intelligent information systems for security of citizens in urban environments are based on different methods most of them using the advanced in multimedia information processing and communications. From the available set of multimedia information visual and audio information contain the predominant part useful for observation, searching and detection of objects, persons or other areas of interest in real work of the intelligent information systems for security of citizens. In this article are proposed the structure of a suitable multimedia surveillance system as a mobile station in a global intelligent information system for security of citizens. Also the advanced methods for processing of audio-visual information in this multimedia station are proposed and tested modeling some real situations of detecting and tracking the objects, like persons, car etc.

Key-Words: - multimedia; surveillance; audio-visual systems; objects tracking; mobile robots

1 Introduction

The mobile multimedia surveillance station is a basic part of the intelligent information systems for security of citizens in urban environments. There are many known surveillance system [1, 2] which are using video or additional audio sensors to collect and to process video and audio information when observing an area of interest. Many of these systems are very simple and without any elements of intelligence. They only collected and stored video and audio information and have the possibilities only to detect and to alarm the changes in the received sequences of images or the amplitude changes in the input sound signals [3, 4]. The goal of this article is first to propose an appropriate and more effective structure of an intelligent multimedia surveillance system realized as a mobile station in a global intelligent information system for security of citizens in urban environments. The second goal is to propose and test the advanced methods and algorithms for joined audio-visual objects tracking. The third goal is to propose the implementation of the tested algorithms in an intelligent mobile robot audio-visual system as a new part of an intelligent information system for security of citizens in urban environments.

2 The structure of an intelligent Multimedia Surveillance System

A. General View of the Intelligent Multimedia Surveillance Mobile Station

The proposed structure of an intelligent multimedia surveillance system is presented in the Fig.1.

- It consists from four parts:
  - video camera modul;
  - audio modul;
  - audio visual mobile robot module;
  - desk computer and communication module.

The video camera module is equipped with the two type of video cameras listed as Video Camera in Fig.1. The first type can be considered as a stationary video sensor in the multimedia surveillance mobile station, which can observe the outdoor area outside the mobile station. The real types of this stationary video camera can be chosen from each type of the standard television cameras, but it is recommended to prefer the digital video cameras, with some of the wide spread digital cable interfaces like IEEE1394 or Fire Wire, DVI or HDMI.

Other recommendation for the proposed structure of the video camera module is the possibility to use more flexible type of a video camera equipped with a pan-tilt device to control and direct the camera more precise in the direction of the tracking objects or persons. Also the application of a stereo video camera can be recommended when it is necessary to observe and analyses the depth in a 3D visual area of observation.
In the proposed Video Camera Module it is added Wireless Camera as a type, which is useful in the cases or situations when it is needed a mobility of a person, held in hand the wireless camera. The images from the wireless camera can be transmitted via Wireless Video Interface connected to the main Desk Computer, shown in Fig.1.

The connection of the ordinary microphone or microphone array to the main Desk Computer can be realized also as wireless, which is shown the Fig.1 with the wireless Microphone Array, Transmitter and Receiver, connected to the Desk Computer.

The next module in the proposed structure of the multimedia surveillance Mobile Station is an advanced proposition to apply the intelligent mobile robots in the area of the modern information systems for secure of citizens in urban environments. The advance in the area of mobile robots have as a result in this time, that the intelligent mobile robot can with an appropriate success perceive, process, recognize and understand audio and video information [5,6].

In the Fig.1 it is proposed to add an Audio Visual Mobile Robot Module. The main parts of this module are the audio and visual sensors placed on the mobile robot platform. The type of the visual sensor chosen to be a convenient flexible Pan-tilt stereo Camera connected possibly with a digital interface to the mobile Robot Computer. Audio sensors are presented in the Mobile Robot Module in Fig.1 as a Microphone Array, connected with a Microphone Array Interface to the Mobile Robot Computer.

B. General Description of Algorithm for Objects and Persons Detecting and Tracking from Visual Information Processing with Video Camera Module in Intelligent Multimedia Surveillance Mobile Station

The first part of the proposed structure of an intelligent multimedia surveillance system, presented in the Fig.1, is Video Camera Module. The received visual information is continuously transmits via cable or wireless connection to the Desk Computer, placed in the proposed Multimedia Surveillance Mobile Station. The goal of Video Camera Module is to receive with one of the Video Cameras shown in Fig.1 the visual information from the observed area of interest and to prepare some of the suitable operations, processing the received visual information for detecting and tracking the existing objects or persons in the images. It is proposed in this article an algorithm for objects and persons detecting and tracking from visual information, processed with Video Camera Module in the intelligent multimedia surveillance mobile station. The proposed algorithm consist from the following parts:

- image frame pre-filtering;
- moving detection;
- objects or persons separation;
- object or persons tracking.
3 Detailed Descriptions of the Parts of the Proposed Algorithm for Detecting and Tracking Objects with Video Camera Module in the Intelligent Multimedia Surveillance System

A. Image Frame Pre-Filtering in the Algorithm for Objects and Persons Detecting and Tracking from Visual Information Processing with Video Camera Module in Intelligent Multimedia Surveillance Mobile Station

The algorithm for objects and persons detecting and tracking can work correctly if in the image there are not present the image intensity changes caused from the fast illumination drifts, noise etc. These changes can lead to some errors of false moving object detection and tracking. To overcome these negative effects it is proposed first to perform a pre-filtering of each frame IFr(n) using a filter structure presented on the Fig.2

\[
IFr(n) = iIFr(n) + rIFr(n) \quad (1)
\]

shown that components \(iIFr(n)\) for illumination and \(rIFr(n)\) for reflection are mixed. They can be separated with a Low Pass Filter and a block of addition \(\Sigma\), using the property of the luminance component \(iIFr(n)\) to contain only the low frequency spectral component:

\[
rIFr(n) = LIFr(n) - iIFr(n). \quad (3)
\]

The separated illumination \(iIFr(n)\) and reflection \(rIFr(n)\) are put under inverse Exponential Transform to output them as two separate components \(iFr(n)\) and \(rFr(n)\). Only the reflected part \(rFr(n)\) carry the information for the moving objects or persons in the images. Only this part \(rFr(n)\) is applied as input of the motion detection and tracking algorithm

B. Moving Objects Detection in the Algorithm for Objects and Persons Detecting and Tracking from Visual Information Processing with Video Camera Module in Intelligent Multimedia Surveillance Mobile Station

The moving objects detection algorithm is presented in the Fig.3

It must detect and separate the moving objects from the stationary ones. The moving objects can be persons or other kind of objects, which can be treated as interesting places in the area of observation with the proposed intelligent multimedia mobile station. The algorithm of moving objects or persons detecting and then tracking consist of two or three main parts:

- motion detection in a picture;
- object and human body separation;

separated object or person tracking in a sequence of image frames.
The first step - motion detection have an important meaning for whole tracking process. The input frame sequence is composed of the image frames \( IFr(n) \) and “n” signify the current image frame in this sequence. It is arranged and shown in Fig. 3 an One Frame Time delay, which means to store the previous image frame \( IFr(n-1) \) in a frame memory. The main step in motion detection algorithm is to calculate Image Absolute Difference \( IAD \) as:

\[
IAD = |IFr(n) - IFr(n-1)|
\]  

where:

\( IFr(n), IFr(n-1) \) and \( IAD \) are two dimensional image matrix of current, previous and absolute difference frames, respectively.

The subtraction is made between each of the corresponding pixels of the image matrices of the current and previous frame. It is possible to suppose that the space and temporal object or person movements are in the range of low space and temporal frequencies, so the next step in the algorithm presented on the Fig.3 is mark as Low Pass Filter. It is designed as a sliding local windows moving in the Image Absolute Difference frame IAP. In each position of the local window is calculate a value of absolute difference \( VAD(i,j) \):

\[
VAD(i, j) = \sum_{k=1}^{M} \sum_{l=1}^{N} IAD(i-k, j-l)
\]

where:

\( i,j \) are the row and column values or co-ordinates of the current center position of the local window and \( i = 0,1,...,N_x - 1; \ j = 0,1,...,N_y - 1; \)

\( N_x \) and \( N_y \) - the horizontal and vertical image size;

\( k = 1,2,...,M \) and \( l = 1,2,...,N \) - the current offset of the local window elements in respect of their center position;

\( M \) and \( N \) - the local window vertical and horizontal size.

The calculated values of the absolute difference matrix are transmitted to the next step of the algorithm, where it is made a decision of these values, with a chosen threshold \( \theta \), if they belong or not to the moving parts or objects in the processed visual information:

\[
BIM(i, j) = \begin{cases} 
0 & \text{if } VAD < \theta \\
1 & \text{if } VAD \geq \theta 
\end{cases}
\]

The result is a binary image mask \( BIM \) containing values “0” for pixels belonging to the static image regions and values “1” for pixels belonging to the moving image regions. From the binary mask it is possible simple to calculate the speaker position \((x_{sp},y_{sp})\) for example as centre of gravity and then in the next step of the proposed algorithm to track only the changes of the co-ordinates of this centre of gravity.

C. Objects or Persons Separation and Tracking

The results from the previous step of the algorithm is a binary image mask \( BIM \) contain information as values “1” for the pixels belonging to the moving image regions. The mentioned above simple and often used way to separate and then track the objects or persons in the processed visual information, described above, is to calculate the object or person position \((x_{sp},y_{sp})\). It is proposed a more suitable, precise and effective method to separate and track the objects or persons in the area observed with Video Camera Module. This method is based on the proposition to form a structural shape description of the separated moving regions in the processed image, then to utilize the Distance Transform (DT) in Image Matching operation as it is presented in Fig.4. In the proposed method it is used an Image Training Database to train and test, with the selected sequence of images, the ability of the proposed method first to make Shape Segmentation of the objects, then Shape Explanation to define a Hierarchical Shape Clustering and Tree Structure Description. This description is marked as \( T \) at the output of the Tree Structure Description block, as is shown in Fig.4.

The most important part from Fig.4 is Image Matching, realized with the Distance Transform (DT), which involves two binary images, represented as binary image mask \( BIM \). These two images are template \( T \) and current test image \( I \). The templates \( T \) are collected in Tree Structure Description Stage as a result of Image Training.

From the variety of Distance Transform DT algorithms here, the Chamfer Distance Transform is used:

\[
D_{chamfer,G} = \sqrt{\frac{1}{|T|} \sum_{t \in T} d_i(t)},
\]

where:
Figure 4. The steps of the algorithm for detecting and tracking the existing objects or persons, processing visual information, received with the Video Cameras in Video Camera Module.

The proposed algorithm for objects and persons detecting and tracking from visual information processing with Video Camera Module is used and tested practically in the intelligent multimedia surveillance mobile station. Also this algorithm is implemented in the audio-visual mobile robot system as a new part of an intelligent information system for security of citizens in urban environments.

4 The Proposed Audio-Visual Mobile Robot System as a Part of the Multimedia Surveillance Mobile Station

The mobile robots perform the audio and visual information from the corresponding audio and visual sensors, shown in Fig. 1. Most popular audio and video sensors are microphone arrays and video cameras. They are placed on the mobile robot moving device and collect the sounds and images from the robot space of observation respectively. Sounds are transformed from each of the microphones in the array as sound signals and images from the video camera (or cameras in case of stereo robot vision) as video signals. The main goals in the processing of sound and video signals in the audio-visual mobile robot system are:

- to separate the signal of interest (SOI) from the combination of sound signals, received from microphone array (usually this signal is speech signal from a speaker talking to the robot in area or space of observation;
- to calculate or estimate the sound or speech signal direction of arrival (DOA) (this is a kind of sound localization);
- to separate the visual object from the received pictures (usually the speaker in space of observation);
- to calculate the position of separated visual object from the received pictures;
- finally, to joint the information as co-ordinates of speaker, calculated from audio and video system, and use this joined information for choosing the right direction for the control of the robot moving system.

Some or all of these mentioned tasks can be simulated separately or in combination, but in all cases exist the problem of a fast and simple modification of the characteristics or dimensions of space or area of observation, mobile robot position, speaker places, etc. This mean, that it is necessary to have a tool in computer simulation programs, which gives a visual representation of robot space of observation with the possibility to interactive changes of room dimensions, robot and speakers places, etc. The simple case for design a visual interactive tool in simulation of a robot audio-visual system is to represent the robot space of observation as an empty rectangle, i.e. as a 2D space of observation and then place in this rectangle the initial robot and speaker positions. In general case the robot space or area of observation can be arbitrary, but here it is chosen as an example, to use an empty room as a concrete space of robot observation. It is also accepted to consider the room only in 2D dimensions length “l” and width “w”. This assumption is applicable, because in most the cases the robot movements are on the room floor.

The proposed visual design of 2D space of observation of the robot is presented in Fig.5 as an algorithm of steps, which can be used first for the initial states settings and then as an interactive work in the current simulation of the audio and visual situation in area or space of robot observation. In the Fig.1 are shown only initial steps of the algorithm.

The first step of the algorithm is for the input of the room dimensions length “l” and with “w”, which here are present as variables “rooml” and “roomw” in the algorithm. Then, for the visual representation on a screen, it is necessary to calculate the resolutions Nx and Ny of display in direction x and y (related for example with room length and width respectively). For these calculation, one of these relative and discrete dimensions Nx or Ny can be chosen, for example Nx=1024 to guaranty a good and precise resolution in x direction. Then it is defined the coefficient Kx as a relation between metric and discrete room dimensions:

\[ |T| \text{ denotes the number of features in } T; \]
\[ d_i(t) – \text{ chamfer distance between feature } t \text{ in } T \text{ and the closest feature in } I. \]
\[ K_x = \frac{rooml}{Nx}. \]  

(8)

The equation (8) can be used in the calculation of discrete room dimension \( Ny \) in direction \( y \):

\[ N_y = \frac{roomw}{Kx}. \]  

(9)

The chosen and calculated \( Kx \) from equations (8) and (9) discrete room dimensions \( Nx \) and \( Ny \) can be used for room drawing and displaying as an rectangle. The next step in the algorithm shown in Fig.5 allows interactive to choose the position of the robot in the room space, i.e. in the rectangle, which is already drawn. The interactive chosen robot position is represented in rectangle room space as robot co-ordinates \( xr \) and \( yr \), respectively.

The visualization of the robot place in the area of room rectangle is made with an icon of stylize and simplified image of a robot.

In the similar way the next steps of the algorithm shown in Fig.1, are for an interactive choosing the speaker position in the space of rectangle of robot observation. Then this position is transformed after some calculations to the absolute speaker co-ordinates \( xs \) and \( ys \) in the space of robot observation. This is necessary to display the speaker as an initial place in the space of observation and on an appropriate distance from the robot position.

The results of the visual design algorithm in the steps of initialization are shown in Fig.6.

\[ \text{END} \]

Figure 5. The proposed initial steps of algorithm of simulation the audio and visual situation in area or space of robot observation

Figure 6. Mobile robot space of observation

The initial step is followed by a concrete simulation or testing the work of the audio and video robot system. Each of two robot systems – audio and video execute at the same time different tasks. Audio system performs sound localization and calculates direction of talk to the robot. Video system, also process the images of the room to separate the place of the speaker. The results of the operation of audio and video systems are the direction of arrival and the coordinates of the speaker, which are calculated both from audio and video system:
\[
\theta_d = \frac{\sum_{\theta=1}^{360^\circ} R_y[k(\theta)] \cdot \theta}{\sum_{\theta=1}^{360^\circ} R_y[k(\theta)]}; \\
xs = \frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} a_i}; \quad y_s = \frac{\sum_{i=1}^{n} a_i y_i}{\sum_{i=1}^{n} a_i},
\]

(10)

where

- \( \Theta_d \) is the angle of direction of arrival;
- \( R_y[k(\theta)] \) – cross correlation in received microphone array speech signals in relation to angle \( \Theta = 1^\circ \to 360^\circ \);
- \( xs \) and \( ys \) – speaker coordinates, calculated from visual system;
- \( x_i \) and \( y_i \) – current image co-ordinates;
- \( a_i \) – the brightness value of current image point.

The equations (10) serve as audio and video information for the proposed here visual design and representation of moving robot audio and video system. The interactive part of the visual design with simulation of audio and visual system operation is shown in Fig.7.

5 The Results of the Simulations of Multimedia Surveillance Mobile Station with Audio-Visual Mobile Robot

The results of audio and video system actions are shown with different line styles for clearly separation of real direction (solid line), audio system direction of arrival determination (dashed line) and observation from video system the direction of Speaker1 (dotted line). Also in Fig.8 are shown the values of angle (theta=36\(^\circ\)) and Speaker1 co-ordinates \( xs1 \) and \( ys1 \), calculated from equation (10).

A full sequence of robot movement simulation in direction of the Speaker1 is shown in Fig.9, where it is seen, as a sequence of points, some of the current robot positions, calculated from audio and video system in the movement of the robot to the Speaker1. These points are near as possible to the real direction to the Speaker1.
The presented work of the proposed visual design tool or interface for simulation and testing the ability of audio and video robot system shows, that this visual tool give good possibilities to express all steps: initial, interactive and final of the algorithm, in a very convenient form. Also this visual tool allows analyzing the results for the working, capacity and efficient of the audio and video robot system. In the future works it is expected to extend the view and the capabilities of this tool for visual representation and interactivity in the real time working robot audio and visual system.

Another example with correct results of observations, detecting and tracking of the chosen objects as “human”, “car” or arbitrary “reject” object, with the proposed multimedia mobile station, are presented in the Fig. 10. They show that this structure of multimedia mobile station work correct and can be implemented as a necessary part in the global intelligent information system for security of citizens in urban environments.

Acknowledgment
This work was supported by National Ministry of Science and Education of Bulgaria under Contracts:
- BY-I-302/2007: “Audio-video information and communication system for active surveillance cooperating with a Mobile Security Robot”;
- 4137 МР7: “International information system supporting observation searching and detection for security of citizens in urban environment”.

References