I. INTRODUCTION

The goal of described project was development of a platform that can be used in applications involving telecommunication, tele-presence, videoconferencing, and information provision to humans in form of multimedia content (audio-visual information). As a result of this project, the robot called NITRO (Novi Sad Institute of Technology Robot) was developed.

In this paper we discuss the incorporation of advanced videophone features into a mobile robot having multimedia capabilities, adjusted to human interaction [1]. Section II gives an overview of the implemented advanced videophone features. Section III depicts the overall software platform. The conclusion is given in section IV.

II. SYSTEM FEATURES

Fig. 1 shows the hardware architecture of the target platform.

At this stage of development, NITRO is able to establish a network connection with remote operator using wireless network. Through its desktop computer attached to wired LAN, the remote operator has full control over the robot within the same network: it can move it using keyboard or joystick; it can start playback of predefined multimedia content (e.g. movie, company presentation); it receives live video and audio from the robot together with the robot status (sensor readings, speed and presentation playback status); using keyboard or joystick it can move the camera in order to get a better overview of the environment; it can choose a target to lock the camera on. Also, the operator can initiate a videophone conversation with a person near the robot.

III. SYSTEM SOFTWARE

The system software is organized as shown in Fig. 3.

A. Network properties

In order to make the system as much mobile as possible, wireless network is used for connection with the remote operator. Providing bandwidth up to 11Mbps at ranges up to 100 meters, the IEEE 802.11b standard was chosen [2].

Using this standard makes the wireless nature of the connection transparent for the remote operator. The mobile unit and the remote operator are members of same network.

As the effective range of wireless access points is up to 30 meters in closed space, a building (or a floor) can be covered by using multiple access points. A mobile network client can leave the area covered by one access point and to
enter an area of another one. In order to preserve continuous connection with the remote party, the mobile client has to seamlessly switch from one access point to another one. This is called roaming. It is defined by IEEE 802.11f standard called Inter-Access-Point Protocol (IAPP). A prerequisite for roaming in a wireless network is to support for IAPP by the wireless access points. In appropriately configured network the mobile client will roam from one access point to one with stronger signal.

![Network structure provided by IEEE 802.11b](image)

Fig. 4 Network structure provided by IEEE 802.11b

As a communication protocol UDP from the TCP/IP protocol stack is used. The commands from the remote operator are transmitted as UDP datagrams.

The remote commands are extracted from UDP packets and they are routed to the appropriate subsystem. The motion subsystem handles the motion related commands, while the multimedia subsystem handles the multimedia and videophone related commands.

The motion subsystem performs motion related tasks like motor control, sensor control and semi-autonomous control.

The multimedia subsystem implements the videophone functions. Its structure is shown in Fig. 5.

![Multimedia subsystem software structure](image)

Fig. 5 Multimedia subsystem software structure

B. Streaming protocol

The network communication is performed at UDP level – the compressed multimedia content is divided into UDP packets of up to 1500 bytes, which are transmitted to the other side. Due to the characteristics of multimedia content a dedicated protocol is used, which makes possible the time-line reconstruction of the data stream at the receiver side. The protocol used for streaming is Real Time Protocol (RTP) [3, 4].

C. Audio-video compression

For videophone communication, standard H.263 video codec [5] and G.723.1 audio codec are used [6].

The format of the video stream is QCIF (176x144 pixels), up to 25 fps, requiring up to 128 kbps. The attributes that influence the video stream bandwidth are the frame rate and quantization level. As network conditions change, the attributes are appropriately changed. The achieved frame rate is typically around 15 fps.

The G.723.1 codec compresses the mono sound sampled at 8 kHz to a data stream of 6.3 kbps. As it compresses one block of 240 samples to 24 bytes, the overhead caused by network header (at least 28 bytes for UDP packets, additional 12 bytes for RTP) becomes significant, consuming 15.3 kbps of bandwidth.

D. Synchronized, adaptive playout

Before the compressed multimedia streams reach the receiver, they are transmitted though the network as independent streams. After decoding, the timings of streams are ruined. The task of the playout module is to reconstruct the correct intra-media and inter-media timings.

The network can influence the transmission in several ways: packets can be reordered and they can be significantly delayed or even lost, especially in wireless networks [7].

To achieve quality videophone communication, several requirements has to be fulfilled:

- The timings within one data stream have to be reconstructed.
- The audio and video streams have to be played back synchronized. This problem is known as lip-synchronization.
- The end-to-end latency must be less than 400 ms for smooth human conversation.

The developed playout algorithm aims to minimize the latency by providing best playout quality. The playout quality is indicated by the number of the missing packets.

Beside the issues addressed above, the algorithm also compensates different clocks at end-points (clock-skew) [8], dynamically changing network conditions (jitter compensation), and sudden jumps in network delay (transit time spikes) [9].

E. Visual tracking

The video processing includes algorithms for face detection, face recognition and face tracking.

![Face detection phase](image)

Fig. 6 Face detection phase

Face detection is first phase in face tracking process. Face is detected by using stored color histogram model. Detector searches ellipse [10] surrounds area in picture containing color information that best matches stored color histogram model (see Fig. 6). Search for face is performed in area around central point specified by the user. Algorithmic tool used in this process is also used in face tracking.

The face recognition function is based on eigenface approach derived from principal component analysis [11].
Selected image is pre-processed by simple normalization and transferred into eigenface space. Eigenfaces are characteristic features of the faces extracted as most dominant from face-space. Because we use only a dominant part of the eigenfaces, such description represents an approximation of the original image. Main task of this transformation is significant dimensionality reduction of the face feature vector. Calculating Euclidean distance between eigenface vector of the current image and eigenface vectors of previously trained base of known faces (organized into face classes), it provides a measure of similarity. The average lowest value of this measure, indicate possible recognized face. If the calculated Euclidean distance between current and closest face class is under some threshold value, one can assume that image is not face at all.

The goal of face tracking algorithm is to keep desired person in centre of image. Demand to head tracking subsystem is to operate on live video in QCIF resolution (176 x 144) at 15 fps. Head tracking must be robust i.e. to preserve lock as long as possible on the selected person.

Method for head tracking combines the output of two different modules: one that matches the intensity gradients along the object’s boundary and one that matches the color histogram of the object’s interior (Fig. 7). Since these two modules have roughly orthogonal failure modes, they serve to complement one another. The result is a robust, real-time system that is able to track a person’s head with enough accuracy to automatically control the camera’s pan, tilt, and zoom in order to keep the person centered in the field of view at a desired size. The algorithm is insensitive to out-of-plane rotation, tilting, severe but brief occlusion, arbitrary camera movement, and multiple moving people in the background [12].

When the remote (far-end) participant is talking, the voice is recorded using far-end microphone, and it is transmitted to the near-end, where it is amplified and sent out to the loudspeakers. The near-end microphone also receives this signal directly from the loudspeakers, but also as several reflected paths in the room. If the system does not contain an echo canceller, the ‘far-end’ speaker will hear several delayed versions of its own voice.

The echo canceling algorithm consists of following algorithm blocks: delay estimation (DE), voice activity detector (VAD), double talk detector (DTD), finite impulse response (FIR) filter with Fast Least Mean Square (FLMS) algorithm, and post processing algorithm (PP) (see Fig. 8). The VAD algorithm analyzes the input signal and determines if it contains human speech or not. The information about voice activity is input to DE and FLMS software blocks. Only if there is human voice in the input signal, delay estimation will be done, and FLMS block should adapt the FIR filter coefficients.

Delay estimation block estimates the delay between the transmitted and received signal.

The double talk detection algorithm determines if the there is speech signal present in both received and transmitted signals simultaneously. If double talk is detected, adaptation procedure of FIR filter is disabled.

The FLMS software block implements fast least mean square algorithm for FIR filter coefficients adaptation [14]. The FIR filter is used to model the echo behavior based on transmitted audio. The FLMS algorithm operates in frequency domain. The PP algorithm aims to eliminate residual echo. It uses as input signal the output from FLMS block, containing echo that FLMS block did not manage to suppress. It operates in frequency domain and it uses sub-band approach. The averaging recursive filters of first order are dedicated to each sub-band. The estimated echo is subtracted from the input signal.

\[ \text{Out}(k) = \text{U}(k) - E(k) \]

**Fig. 7 Head search processing flow**

Calculated motion vector is used as input for calculation of parameters of camera movement. PID controller links motion vector with real camera movement vector [13].

**F. Acoustic echo canceling**

The acoustic echo canceling (AEC) subsystem consists of algorithms providing cancellation of an acoustic echo. The need for acoustic echo control arises whenever a loudspeaker and a microphone are placed in position that the signal emitted by the loudspeaker, as well as its reflections at the borders of the enclosure (room), are picked up by the microphone. If the microphone signal is fed back to the loudspeaker in some way, the electro-acoustic system may become unstable and produce howling.

\[ D(k) = \text{Y}(k) - E(k) \]

\[ E(k) = \text{D}(k) + \text{Out}(k) \]

**Fig. 8 Structure of the echo canceller**

Inputs of AEC are transmitted signal (signal that is to be emitted by the speakers), and signal recorded by the microphone. Output from the AEC system is the useful content calculated by subtracting the estimated echo from the recorded sound.

The VAD algorithm analyzes the input signal and determines if it contains human speech or not. The information about voice activity is input to DE and FLMS software blocks. Only if there is human voice in the input signal, delay estimation will be done, and FLMS block should adapt the FIR filter coefficients.

Delay estimation block estimates the delay between the transmitted and received signal.

The double talk detection algorithm determines if the there is speech signal present in both received and transmitted signals simultaneously. If double talk is detected, adaptation procedure of FIR filter is disabled.

The FLMS software block implements fast least mean square algorithm for FIR filter coefficients adaptation [14]. The FIR filter is used to model the echo behavior based on transmitted audio. The FLMS algorithm operates in frequency domain.

The PP algorithm aims to eliminate residual echo. It uses as input signal the output from FLMS block, containing echo that FLMS block did not manage to suppress. It operates in frequency domain and it uses sub-band approach. The averaging recursive filters of first order are dedicated to each sub-band. The estimated echo is subtracted from the input signal.
Acknowledgment: We would like to thank Prof. Vladimir Kovačević and Prof. Miodrag Temerinac for their guidelines, suggestions and support during development of this project. Besides, we would like to thank the MicronasNIT people for ideas and valuable comments that have improved this project.

Abstract: In this paper, a multimedia features integrated into remotely controlled mobile robot, are described. Building such a complex system requires knowledge from several scientific fields: mechanics, electronics, hardware, software, telecommunications and even design. As the developed robot is intended to tightly cooperate with humans, the ergonomic criteria had high priority. The integration of building blocks is always followed by some adaptation for the specific application, requiring additional research and improvements of the existing solutions. Thus, during development, the original concept of the robot evolved as new requirement were recognized and as they are added in form of new features. Finally, the developed robot is able to move according to remote commands while retaining its integrity; it can play back various multimedia content; it provides a bidirectional videophone link between the remote operator and the person near the robot (partner), which offers new, exciting features like visual recognition and tracking of the partner, and advanced acoustic echo suppression algorithm.

REFERENCES


