Toward Ubiquitous Video-based Cyber-Physical Systems

Guoliang Xing[1]; Weijia Jia[2]; Yufei Du[2]; Posco Tso[2]; Mo Sha[2]; Xue Liu[3]
[2]Department of Computer Science and Engineering, Michigan State University
[1]Department of Computer Science, City University of Hong Kong
[3]School of Computer Science, McGill University

***Abstract*— Cyber-Physical Systems (CPS) is a new generation of engineered systems that integrate physical systems with the capability of networked computing and control. Real-time video capture and communication is expected to be an important function in many Cyber-Physical Systems that involve camera-equipped mobile phones. In this paper, we present AnySense, a network architecture that supports video communication between 3G phones and Internet hosts in Cyber-Physical Systems. AnySense implements transcoding of video streams between the Internet and circuit-switched 3G cellular networks, and is transparent to 3G service providers. AnySense can support a class of ubiquitous Cyber-Physical Systems that require video-based information collection and sharing. A prototype of AnySense has been built and a video demo is available at http://www.anyserver.org/.**

I. Introduction

Cyber-Physical Systems (CPS) has emerged as a new generation of engineered systems that seamlessly integrate physical systems with the capability of networked computing and control, i.e. the cyber systems. CPS often requires information sharing and coordination between physical processes, the Internet and human. The advances of mobile communications and embedded sensing have made it possible to build large-scale cyber-physical systems composed of mobile phones that collect information about the physical environments at anytime and from anywhere.

We envision the popularity of a class of video-based sensing applications [5], [1], [2] in which camera-equipped mobile phones are used to capture, send and receive *real-time* videos for CPS. For instance, a mobile phone user may record (physical and social) events occurring in his proximity and send to friends or broadcast through the Internet in real-time. For another example, voluntary mobile phone users may collaborate to collect and broadcast real-time video of city-wide traffic condition during rush hour. In addition to video capture, mobile phones can also be used for receiving videos from cameras connected to the Internet. For instance, a mobile phone user may want to receive real-time video sent from the surveillance cameras installed in his/her home once an abnormal event is detected. In the aforementioned applications, real-time video provides much richer information about the events of interest than other types of data formats such as text messages or static images.

Compared with existing applications on hand-held devices like video downloading and viewing, video-based cyber-physical systems introduce several new challenges. First, in order to capture and publish the video of unpredictable physical and social events, mobile phones must have ubiquitous broadband network access, which puts the solutions that rely on short-range wireless networks into question. For instance, although wireless LANs can provide high bandwidth for the smart phones equipped with WiFi interfaces, they usually only cover a small portion of a metropolitan area due to the short range. WiMAX and wireless mesh networks are two emerging technologies that have the potential to provide wide-area high-speed Internet access. However, their deployments are still in a nascent stage. Second, many video-based sensing applications (e.g., instant news coverage) are expected to involve unplanned information collection and substantial numbers of participants. As a result, video capture and communication should be implemented by "out-of-the-box" functions (e.g., video calls) on off-the-shelf mobile phones. For instance, it is undesirable to require all the mobile phones in an application to install additional software or support a special functionality that is only available on the mobile phones from particular manufacturers.

We develop a video communication architecture based on 3rd-generation (3G) mobile technology, called AnySense, to support video-based cyber-physical systems. 3G networks are capable of providing both high-speed

1-4244-2384-2/08/$20.00

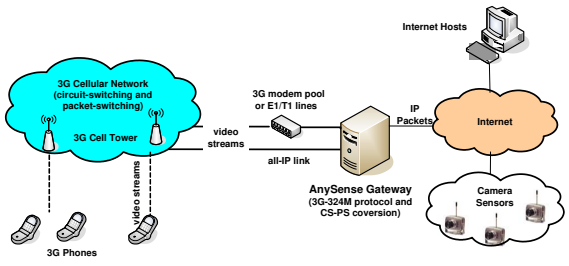

Fig. 1. Architecture of AnySense

data and voice/video call services. The number of 3G cellular network deployments has been rapidly increasing worldwide. Video communication between 3G phones and Internet hosts can be achieved by all IP links. However, real-time video streams often suffer from high jitters when transmitted over 3G IP links due to poor link quality and variable bandwidth [9]. As a result, complex software QoS control often needs to be implemented on 3G phones to support IP video calls.

The solution of AnySense to the above issue is to seamlessly bridge circuit-switched (CS) 3G networks with the packet-switched (PS) Internet. Video data from mobile phones will be carried by CS wireless links with guaranteed bandwidth and built-in QoS control, and are transcoded into PS video streams at a gateway before being sent to Internet hosts. This approach has several advantages compared to the all IP solution. First, due to the built-in CS video call support, no change or installation of new software is needed for a 3G phone to communicate with Internet hosts. This feature is particularly important for ubiquitous video collection and communication, e.g., instant news coverage, which may involve a large number of participants that carry off-the-shelf 3G phones. Second, connections over CS have built-in support for bandwidth reservation and QoS control, which is key to achieving optimized video quality on resource-limited mobile phones.

AnySense supports a number of video-based cyber-physical systems that involve 3G phone users. We now describe a real-time video surveillance application that can be built on top of AnySense. In this scenario, camera sensors are installed in the home/office of a 3G user for security surveillance. We assume that a camera sensor is capable of analyzing the videos captured (e.g., detecting human motions). When a camera sensor detects an event, it transmits the real-time video (through a wired/wireless LAN) to a base station, e.g., a PC connected to the Internet. The PC then initiates a UDP connection with the AnySense gateway and sends the real-time video recorded by the camera. The gateway dials the number of the 3G user, establishes a video call connection with the 3G phone, and sends the video stream through 3G circuit-switched cellular network.

II. System Architecture

Fig. 1 illustrates the architecture of AnySense. The core component of AnySense is the gateway that interconnects the 3G cellular network and the Internet. For instance, when a 3G phone sends a video stream to an Internet host, the 3G phone initiates a video call to the gateway and the video stream is then transmitted over the 3G CS network. The gateway then converts the stream to IP packets and sends to the host using TCP, UDP or RTP. The gateway converts IP packets to 3G CS network in a Similar way. With an IP address and a phone number, the gateway essentially serves as the proxy for both the 3G phone and the Internet host. We note that the operation of the gateway is transparent to 3G service providers. Furthermore, a 3G phone handles video streams from an Internet host in the same way it handles video calls from 3G phones. Therefore, our architecture supports the video communication between Internet hosts and any off-the-shelf commercial 3G phones.

The gateway implements three important functionality. First, it implements 3G video circuit transmission and Internet signaling protocols including 3G-324M [10] and Session Initiation Protocol (SIP) [3] stacks. Second, it performs conversion between 3G circuit switching and IP packet switching. Third, it can automatically switch between CS and PS channels depending on the video quality that can be achieved by each channel.

3G-324M is the International Telecommunication Union (ITU) standard for real-time multimedia services over circuit-switched cellular networks. 3G-324M includes an umbrella protocol, H.324M [4], and several core protocols such as control protocol H.245 and (de)multiplexing protocol H.223. Several key features are defined by 3G-324M to support real-time streaming over circuit-switched networks, including fixed communication delays and low-overhead codecs. H.245 is used by 3G-324M to perform call connection setup and tearing down, capability and codec negotiation, etc. Data (de)multiplexing is performed by H.223. At the multiplexing stage, control and multimedia data (e.g., voice and video) are multiplexed into a bit stream with appropriate delimiters and output to the physical layer (e.g., air interface). At the de-multiplexing stage, multimedia data are extracted from the bit stream received from the physical layer.

We now briefly describe the process of converting CS video streams to IP packets. The conversion from IP packets to CS video is similar. First, an incoming video call is accepted by the modem pool or T1/E1 interfaces on the gateway. A T1/E1 link is a fiber optic line that can handle 24 call connections. Second, after the connection is established, data breakdown and re-encapsulation are performed on the video streams sent from a 3G phone by H.245 and H.223 protocols. Third, voice and video data are transcoded through efficient media codec such as H.263 and H.264. Finally, Session initialization protocol (SIP) manages the multimedia sessions among the gateway and Internet hosts. SIP has been widely used as signaling protocol for Voice over IP (VoIP). Therefore, AnySense gateway supports the communication between 3G phones and any SIP based VoIP softphones. Depending on network environment, TCP, UDP or RTP is used to transmit data between the gateway and Internet hosts.

Another important functionality implemented by the AnySense gateway is the quality-driven channel switching between CS and PS for mobile phones. Specifically, when a mobile phone is receiving a video stream, the gateway monitors the network condition and estimates the quality of experience of the phone user based on empirical models. It then switches between CS and PS to optimize the video quality of the phone. Channel switching has several key advantages. Although the 3G CS channel only has a maximum data rate of 64 Kbps, it has several built-in mechanisms (such as reserved bandwidth) for Quality of Service (QoS) assurance. On the other hand, the PS channel can achieve a much higher data rate (e.g., 384 Kbps at or below pedestrian speeds, and 128 Kbps in moving vehicles), it solely relies on the end systems for QoS assurance. Moreover, the video quality exhibits very different characteristics with respect to network conditions under the two channels. These facts offer a room for optimizing video quality by exploiting the dual channel capability of 3G mobile phones. The details of channel switching are discussed in Section IV.

III. Video Quality Modeling

In this section, we establish an model of video quality for both CS and PS channels of 3G networks. Our model is based on empirical measurements of 18 video clips (which are available at http://www.anyserver.org/). The model correlates a video quality metric called the Perceptual Evaluation of Video Quality (PEVQ) [8] with several network parameters including frame and packet loss rates. In Section IV, we discuss how this model is used for automatic network channel selection.

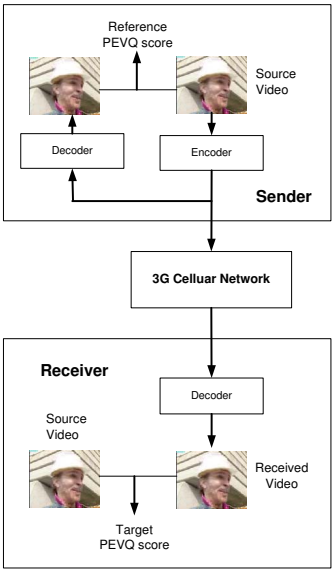

Fig. 2. The procedure of measuring PEVQ score ratio.

A. Modeling Methodology

Our video quality modeling is based on the Perceptual Evaluation of Video Quality (PEVQ) standard from OPTICOM [8]. PEVQ is widely adopted for quantifying the Mean Opinion Score (MOS) of the video quality for IPTV, streaming video, mobile TV and video telephony. A PEVQ implementation compares the quality of two video clips, the original clip and the clip that suffers from quality degradation during network transmission. The result of comparison is quantified by a rational number within 1 (bad) and 5 (excellent). The mechanism of PEVQ is based on modeling the visual behavior of human.

A key challenge to video modeling in 3G networks is to eliminate the impact of video encoders. As the encoding of 3G phones is lossy, the PEVQ score of a video is largely dependent on the specifics of the encoder used by the 3G network. As a result, a encoder-dependent video quality model is not applicable to 3G networks from different carriers. Ideally, a video quality model should only capture the impact of network transmission parameters without the dependence of encoding schemes. To this end, we propose a novel reference-based modeling scheme that can effectively reduce the impact of encoding on video quality quantification.

Our basic idea is illustrated in Fig. III-A. We assume that a 3G mobile phone sends a video clip to another phone using the CS or PS channel. For modeling purpose, both the sender and the receiver store the video clip locally. The sender first encodes the source video to be transmitted. The encoded video is then decoded by a local decoder that is identical to the one used by the receiver. The decoded video and the source video are then input to the PEVQ tool that outputs a score within $[1, 5]$, which is referred to the *reference PEVQ score*. The encoded video is then transmitted through a 3G cellular network. After receiving it, the receiver decodes the video and compare against the source video using the PEVQ tool. The score obtained is referred to as the *target PEVQ score*. We then compute the ratio of the target score to the reference score as the metric that quantifies the impact of network transmission on the quality of video. Intuitively, the reference PEVQ solely captures the loss of encoding because the video has not been transmitted. On the other hand, the target PEVQ score reflects the quality degradation due to the loss of both encoding and network transmission. Therefore, by mitigating the encoding effect, the ratio of the two scores is mainly dependent on the network condition.

We observed that the PEVQ score ratio is mainly affected by two network parameters, namely packet loss rate and frame loss rate. Moreover, the PEVQ score ratio and packet/frame loss rates can be accurately modeled by a quadratic function. Interestingly, the quadratic models for the CS and PS channels bare the same form with only difference in the coefficients. Suppose x, y and z denote packet loss rate, frame loss rate and the PEVQ score, respectively. We define the generic quadratic function $\phi(x, y, z)$ as follows:

$$\begin{aligned}\phi(x,y,z) &= ax^2 + by^2 + cxy + dx + ey + fz^2 \\ &\quad + gxz + hyz + ixyz + jz + k \end{aligned} \tag{1}$$

Our objective is to determine the coefficients $[a, b, \cdots, k]$ in (1) through experiments.

B. Experimental Setup and Results

In the CS experiments, we used two 3G-324M protocol stack capable tools, Dilithium Network Analyzer (DNA) [7] and SkyWalker [6], to carry out the experiments. The Dilithium Network Analyzer is adopted because it is not only the industrial recognized testing tool for 3G video calls but also it generates a variety of logging documents that are critical for us to validate the experiments results. AnySkywalker [6] is a softphone built on top of our own 3G-324M protocol stack, which is used to convert received video streams into AVI files for PEVQ [8] evaluation.

To conduct the experiments, we made 3G video calls from DNA to AnySkywalker three times for every video feed. Several logging files are generated after each run, which contain time stamp, packet sequence number, media content for every packet. The packet delay is ignored as we observed that it is almost fixed in normal 3G video calls because of guaranteed bandwidth. The packet loss is measured by counting the number of missing sequence numbers in log file in every second. In addition to the packet loss, we also count frame loss by comparing the number of frames in the original AVI file with the received one. Finally, the quality of received AVI video clips are scored by the PEVQ tool.

In the PS experiments, we use UDP as the protocol for carrying videos over 3G networks. We extended AnySkywalker to transmit/receive data from UDP sockets. As the CS channel has a maximum data rate of 64 Kbps, we also fixed the sending rate of PS at 64kbps (including 12kbps audio as in 3G CS video calls) for fair comparison. As in CS experiments, we generate log files containing packet time stamp, packet sequence number as well as packet content. Different from the CS experiments, UDP packets arriving at the receiver may be out of order. These packets will be dropped by receiver because packet reordering could lead to significant packet delay and quality degradation of videos.

We used total 18 video clips (which are available at http://www.anyserver.org/) in our experiments. Each video clip is transmitted three times, which results in total 54 runs of measurements. We use simulation annealing to approximate the coefficients in curve fitting. The initial settings of coefficients are randomly chosen. To evaluate the accuracy of our models, we choose a small training set of runs to compute the model coefficients. The obtained models are then used to predict the PEVQ score ratio of the rest of the video clips. We vary the size of training set from 5 to 18. For each setting, we compute the relative error of between the predicted and actual PEVQ score ratios. As the results of CS and PS experiments are similar, we only present the results of CS in this section. Table I lists all the coefficients obtained with different training sets.

Fig. III-B shows the histogram of relative error with different training set sizes. We can see that a majority of errors fall within 0.2. In particular, 85.7% of errors are smaller than 0.2 even when the training set only

Training Set Size	5	10	18
a	0.134483	0.063876	-0.12698
b	283.654	38.8039	-26.7358
c	-9.20878	-74.1357	-10.978
d	-0.09253	0.145062	0.227486
e	-8.52848	46.0365	9.065613
f	-0.05037	-0.02192	0.000439
g	0.033064	-0.0383	0.016822
h	6.226753	-6.51488	1.333796
i	5.659283	17.07271	-0.94643
j	0.146546	0.096076	-0.01926
k	-0.15626	-0.20274	-0.09532

TABLE I

LIST OF COEFFICIENTS WITH DIFFERENT TRAINING SET SIZES.

Training Set Size	5	10	18
Root mean squared error	0.280969	0.168316	0.037804
Sum of squared error	3.868242	1.246539	0.051448
Correlation coefficient	0.943807	0.982038	0.999016
Squared correlation coefficient	0.890771	0.9644	0.998033
Coefficient of determination	0.884861	0.961465	0.997921
Chi-Square	-1.94792	0.463663	0.030797
F-Statistic	383.2885	1137.759	17253.73

TABLE II

ERROR CHARACTERISTICS WITH DIFFERENT TRAINING SET SIZES.

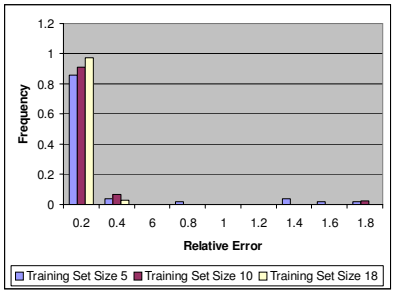

Fig. 3. The histogram of relative error with different training set sizes.

contains 5 runs of experiments, which corresponds to 5/54=9.2% of the total runs. This result demonstrates the effectiveness of our modeling methodology.

Table II shows the error characteristics under various metrics when the training set size varies. We can see that all the metrics drop very quickly when the training set size increases. For instance, the root mean squared error drops from 0.28 to 0.037 when the size of training set increases from 5 to 18.

IV. QUALITY-DRIVEN CHANNEL SWITCHING

In this section, we discuss how the AnySense gateway switches between CS and PS channels based on the video quality models established in Section III. As the CS-PS switching is similar to PS-CS switching, we focus on describing the case of PS-CS switching.

Suppose a Internet host is sending a video clip to a 3G mobile phone. As discussed in Section II, the video clip is first sent to the AnySense gateway, which then forwards the video stream to the phone using UDP. At the same time, the video player client on the phone measures the network condition including packet and frame loss rates. It then sends such information back to the gateway on a control channel. The gateway then predicts the video quality experiencing by the phone using the video model (Eqn. 1). We note that such a feedback mechanism does no require the involvement of phone user. Alternatively, the quality of video may be input by the user. After the gateway receives the feedback, it compares the video quality experienced by the phone user against a present threshold. If the video quality is below the threshold, the gateway initiates a CS connection with the phone and close the current PS connection.

The above channel switching mechanism is particularly advantageous when the mobile phone suffers from highly dynamic network conditions due to mobility. However, online channel switching also incurs considerable bandwidth overhead. The AnySense gateway may also adopt a static channel selection mechanism when the mobility of phone user is not high. Specifically, the gateway and mobile phone conduct a channel negotiation process before video communication. The network conditions including frame and packet loss are assessed for both channels during the process. The gateway then predicts the quality of video based on the video model (Eqn. 1) and then chooses the channel with the better expected video quality. Such a channel switching capability is key to the video communication with stringent QoS requirements.

V. CONCLUSION

In this paper, we present AnySense, a communication architecture for ubiquitous video-based cyber-physical systems. AnySense supports video communication between 3G phones and the Internet hosts. In particular, AnySense implements transcoding of video streams between the Internet and circuit-switched 3G cellular networks and is transparent to 3G service providers. A

class of video-based cyber-physical systems can be built with the support of AnySense. We have built a prototype of AnySense. The details of the AnySense project and a video demo is available at http://www.anyserver.org/.

VI. Acknowledgement

The work described in this paper was partially supported by the City University of Hong Kong under a grant ARD 9668009 and the Research Grants Council of Hong Kong under grants RGC 9041129 and 9041266.

References

[1] Sensorplanet, http://www.sensorplanet.org/.
[2] Mobile landscapes: Graz in real time. http://senseable.mit.edu/projects/graz/.
[3] SIP: Session Initiation Protocol, RFC 3261, 2002.
[4] ITU-T Recommendation H.324: Terminal for low bitrate multimedia communication.
[5] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. Participatory sensing. In *ACM Sensys World Sensor Web Workshop*, 2006.
[6] Weijia Jia, TSO Fung Po, and Lizhuo Zhang. Efficient 3g-324m protocol implementation for low bit rate multi-point video conferencing. In *Journal of Networks*, volume 1, pages 609–616, 2006.
[7] Dilithium Networks. http://www.dilithiumnetworks.com.
[8] OPTICOM. http://www.pevq.org/.
[9] Eli Orr. Understanding the 3G-324M Spec. Online technical tutorial.
[10] 3GPP TR 26.111 V5.0.0. ITU-T Recommendation H.324: Terminal for low bitrate multimedia communication, 2002.