Towards Ultra-Low-Bitrate Video Conferencing Using Facial Landmarks

Paper ID: 392

ABSTRACT

Providing high-quality video conferencing experience over the best-effort Internet and wireless networks is challenging, because 2D videos are bulky. In this paper, we exploit the common structure of conferencing videos for an ultra-low-bitrate video conferencing system. In particular, we design, implement, optimize, and evaluate a video conferencing system, which: (i) extracts facial landmarks, (ii) transmits the selected facial landmarks and 2D images, and (iii) warps the untransmitted 2D images at the receiver. Several optimization techniques are adopted for minimizing the running time and maximizing the video quality, e.g., the image and warping frames are optimally determined based on network conditions and video content. The experiment results from real conferencing videos reveal that our proposed system: (i) outperforms the state-of-the-art x265 by up to 11.05 dB in PSNR (Peak Signal-to-Noise Ratio), (ii) adapts to different video content and network conditions, and (iii) runs in real-time at about 12 frame-per-second.

1. INTRODUCTION

Recent market research [19] depicts that the market share of video conferencing systems is expected to grow from 3.31 billion USD in 2013 to 6.40 billion in 2020, at a compound annual growth rate of 9.3%. Such a high rate of expansion can be attributed to several benefits of video conferencing, such as: (i) lower travel costs on corporates, (ii) less time overhead on employees, (iii) reduced stress and fatigue by avoiding travels, (iv) more effective communications than telephones, and (v) tighter collaborations across multiple offices to cope with globalization. Given that computers and Internet access are more and more affordable, the growth of video conferencing shows no trend of slowing down in the coming years. Nevertheless, providing good video conferencing experience is crucial for its success.

Video conferencing, like many other real-time, interactive multimedia applications, is resource demanding. For example, Skype recommends bitrates between 700 Kbps and 2.5 Mbps for 1280x720 video calls using H.264 codecs [13]. Guaranteeing such a high end-to-end bandwidth requirements is no easy task in the best-effort Internet; and doing so in shared wireless networks, such as WiFi and cellular networks, is even more difficult due to congestion, channel fading, shadowing, and interference.

In this paper, we design an *ultra-low-bitrate high-quality video conferencing system* for commodity computers by analyzing the structures of typical conference video frames and *aggressively* skipping redundant information. By *typical* conference videos, we consider a talking head in a conference room, while our system can be generalized for multiple participants in the same conference room. We make a crucial observation: in these videos, the major movements come from talking heads. To leverage this observation for bitrate reduction, we may: (i) transmit a snapshot image of talking heads at the beginning of each video conferencing session as a reference image, which is referred to as a *base image*, and (ii) describe the talking heads using facial models [17], in order to synthesize facial expressions without sending (bulky) 2D images in all video frames. In particular, we propose such a system (one-way for brevity) in Fig. 1. We divide all video frames from Webcam into two groups: (i) image frames that are transmitted as regular video frames encoded with video codecs and (ii) warped frames that are synthesized using image frames and facial landmarks, such as key feature points on edges of eyes, nose, and mouth. Selected landmarks and image frames are sent to the receiver. The receiver reconstructs the warped frames and sequentially plays all video frames. Fig. 1 also presents sample reconstructed frames from our proposed system and those from a conventional (image-based) video codec at 25 Kbps. The blocking features of the sample frames from the image-based video codec are clear, leading to degraded video conferencing quality. More experiment results are given in Sec. 4.

We emphasize that our goal is very aggressive, as we aim to provide *acceptable* video conferencing quality at 30 Kbps, which is even lower than some audio codecs, such as G.711. To cope with this challenge, we carefully design and implement the individual components in the proposed system. The crux of the whole system is the *frame type selector*, which analyzes the expected video quality of sending the current video frame as: (i) an image frame or (ii) a warped frame under the bandwidth constraints. The frame type selector then makes the decision based on the analysis results, so as to maximize the overall video quality and thus the conferencing experience. We implement the proposed system and conduct experiments with real conferencing videos from several subjects. The experiment results show the merits of our system, for example, it: (i) outperforms the state-of-the-art x265 by up to 11.05 dB in PSNR (Peak Signal-to-Noise Ratio), (ii) adapts to different video content and network conditions, and (iii) runs in real-time at about 12 fps (frame-per-second).

2. RELATED WORK

Qi et al. [16] propose to skip transmitting some video frames at the sender, and employ 2D interpolation to synthesize the skipped video frames at the receiver. Different from our system, their solution ignores the fact that facial expressions of the talking head is the most crucial content in video conferences, and may lose the opportunity to exploit the redundancy among faces in adjacent frames. Facial models may be used to convey more facial details, e.g., Allen et al. [2] extract facial model parameters for higher compression ratios, and Zeng et al. [23] propose a solution to emphasize the appearance of mouths and eyes during video conferences. These stud-

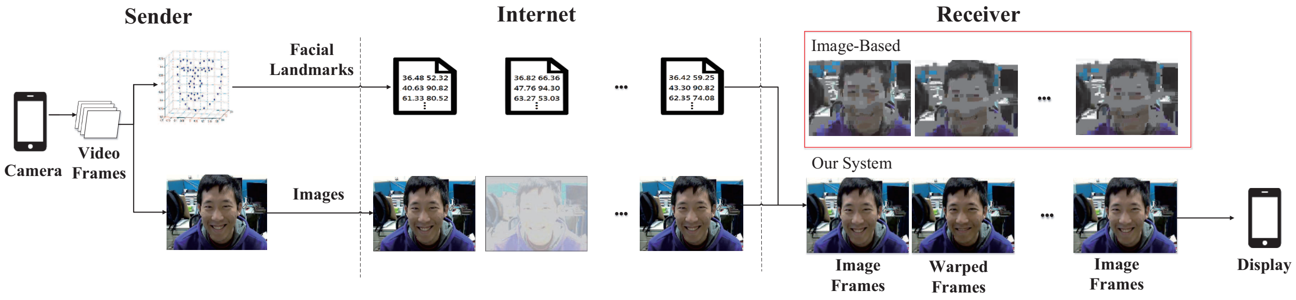


Figure 1: Illustrations of our low-bitrate video conferencing system. Only one-way video streaming is shown for brevity. Sample reconstructed video frames from our and image-based systems are given.

ies [2,23] transmit potentially duplicated texture parameters across adjacent video frames. In contrast, we aggressively avoid sending repeated information by using image frames and landmark coordinates. An early study [10] transmits images from three difference angles of a face, and uses manually-selected landmark coordinates for frame interpolation. The authors find the facial expressions are not clear, and apply Principle Component Analysis (PCA) to transmit the regions of eyes and mouth. Our work also leverages base images similar to [10]; however, their system requires human interventions, and thus is not suitable to interactive video conferences. MPEG-4 [14] provides coding tools for face animation. In particular, the high-level expressions, like, joy, anger, and sadness, can be encoded with a given facial model. However, extracting these high-level expressions is time-consuming (3.72 fps at 352x288 on smartphones [18]), rendering it not suitable for real-time video conferences. MPEG-4 also supports low-level expressions, which are encoded using six parameters, such as the distance between upper and lower eyelids. Six parameters are too few, and lead to degraded facial expressions, as reported in Zeng et al. [23]. Compared to the high- and low-level MPEG-4 coding tools, our proposed system runs faster time and encodes more facial details, respectively.

3. FACIAL LANDMARK-BASED VIDEO CONFERENCING SYSTEM

3.1 Overview

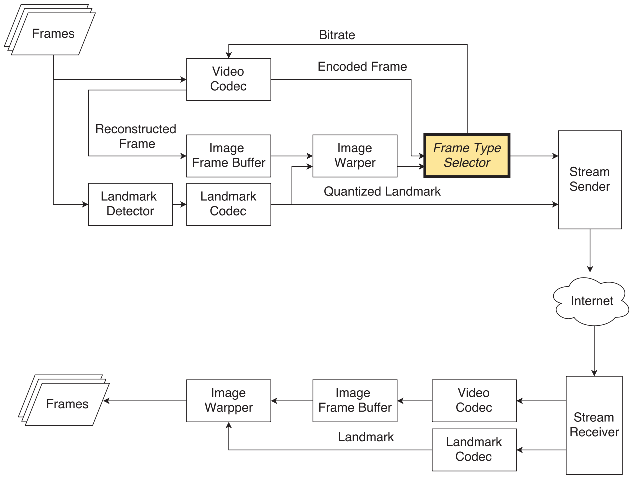


Figure 2: The architecture of our proposed system.

Fig. 2 presents the architecture of our ultra-low-bitrate video conferencing system. Different from conventional video conferencing systems, the proposed system consists of five unique components: (i) *Landmark Detector*, (ii) *Video Codec*, (iii) *Landmark Codec*, (iv) *Image Warper*, and (v) *Frame Type Selector*. We assume the base image is sent to the receiver beforehand, which is used to bootstrap frame warping. The interactions among those components are as follows. The original video frames are captured by a Webcam, and sent to the landmark detector, which outputs the coordinates of detected landmarks. Afterwards, the landmark codec quantizes and compresses the landmarks and transmits them to the image warper. The image warper synthesizes the latest image frame based on the quantized landmark and passes the image to the frame type selector. The frame type selector compares the video quality between the warped frame and the image frame to decide the frame type. Following the decision, the image frames and facial landmarks are transmitted as needed. At the receiver side, the image frames are decoded by the video codec and the warped frames are reconstructed by the uncompressed landmarks and the last received image frame. Among the five components, the frame type selector hosts the intelligence of maximizing video quality. We present the designs of the four other components in Sec. 3.2. and the details on frame type selector in Sec. 3.3.

3.2 Component Designs (Except Frame Type Selector)

Landmark Detector. Facial models can be roughly classified into 3D [3, 5] and 2D [7, 8] ones. We adopt 2D facial models since the illumination and talking head features are rather static in video conferencing. In particular, we consider two most popular 2D facial models: Active Appearance Model [7] (AAM) and Constrained Local Model [8] (CLM). The goal of AAM is to fit a static shape and appearance model to a new image. Through iterations, AAM fits the models and computes the coordinates of the landmarks. In contrast, CLM builds shape models by labeled training data. Generally, AAM preforms better on face alignment and rotated faces, and CLM provides more delicate changes on different expressions. Considering the properties of conferencing videos (more facial expressions, fewer changes on face orientations), we choose CLM with 68 landmarks [4] for better video quality.

Image Warper. Image Warper synthesizes frames based on landmarks and prior image frames. We first use Delaunay Triangulation algorithm [11] to divide faces (excluding the background region) into triangles based on the landmarks. Then, an affine transformation is applied to those triangles. This process can be viewed as if those triangles are projected onto new surfaces pixel by pixel. Moreover, the affine function is ideal for image mapping when the transformed regions are small due to its geometric properties. In our application, the affine transformation not only solves the deformation problem on facial regions, but also incurs lower computational complexity because of its linear property.

Video Codec. We adopt the state-of-the-art H.265 codec, which achieves higher coding efficiency than earlier codecs, such as H.264 and MPEG-4. For example, Ohm et al. [12] report that H.265 achieves 50% bitrate reduction compared to H.264 in objective tests. Their subjective tests show even larger gaps. If codec availability is a concern, any other low-complexity 2D video codec can also be used in our system.

Landmark Codec. Since coordinates of the same landmark impose temporal redundancy across neighboring frames, we encode the landmark coordinates in *deltas* instead of raw numbers, as detailed below. First, we send the number of landmarks in each video frame, which is an 8-bit unsigned integer. This is followed by a series of video frames. Each frame starts with a timestamp as a 16-bit unsigned integer, followed by a series of landmarks. The landmarks are first normalized to the width/height of the video resolution (i.e., between 0 and 100%), and then quantized into the unit of ten thousands. For the first frame, we store each landmark as a 14-bit integer. The landmarks in the following frames are stored in deltas, which can be either a *short* or *long* delta, as indicated by a flag bit. Using real conferencing videos (detailed in Sec. 4), we analyze the distributions of landmark deltas. We find that 4-bit short deltas cover 76.2% of all landmark deltas, and 7-bit long deltas cover 24.5% of them. Since there are only 0.3% remaining landmark deltas, we decide to encode them in 7-bit, and carry over the residues to the next frame in the worst case. With the proposed representation format, we reduce the size of landmark coordinates by about 30%. The resulting landmark representations are compressed by the 7z [1] compression algorithm before being transmitted.

3.3 Maximizing Video Quality Using Frame Type Selector

We design the frame type selector to be content dependent and network adaptive. It dynamically instructs: (i) the sender to transmit an image frame or (ii) the receiver to synthesize a warped frame, so as to maximize the video quality of every single frame. In particular, the sender keeps track of the available bitrate, or *bit budget*, of each video frame. The sender encodes the current video frame using the bit budget and also simulates the warping procedure at the receiver side. The frame type selector then chooses the type (image or warped frame) that leads to higher video quality. Notice that we only transmit the landmarks whenever needed. In particular, warped frames only requires the landmarks of the latest image frame and the current frame. By deferring the transmission of the landmarks of image frames, we never transmit the landmarks that will not be used. We note that our frame type selector implicitly determines the frequency of sending an image frame based on the video content characteristics and network resource availability, which is the core research problem to optimize the proposed ultra-low-bitrate video conferencing system.

4. EXPERIMENTS

4.1 Setup

We have implemented our proposed video conferencing system using CLM [4] and OpenCV [6] libraries, and in C++. Our system has five components: (i) *landmark detector*, which analyzes video frames and generates the landmarks for each video frame, (ii) *frame type selector*, for selecting whether it is an image frame or a warped frame under different total bitrate constraints, (iii) *encoder*, which compacts and compresses landmarks, and invokes x265 [22] to encode images frames according to frame type and the per-frame bitrate constraints, (iv) *decoder*, which decodes the images and landmarks, and (v) *warper*, which generates intermediate synthesized images with OpenCV [6]. The base images are compressed into JPEG, and the x265 configurations are ultrafast preset, zero latency tuning, and IPPP $\cdots$ structure, if not otherwise specified.

We recruit nine subjects in our university, and record nine videos at 1280x720 using commodity Webcams and computers for our experiments[1]. When recording the videos, we ask the subjects to talk as if they are in video conferences. Each video has 300 frames and lasts for 10 seconds. We adopt the following performance metrics in our experiments.

- **Video quality:** the video quality in PSNR and SSIM (Structured Similarity Index [21]), computed by comparing the original video at the sender against the reconstructed one at the receiver.
- **Warping ratio:** the ratio between the number of the warped frames and the total frames.
- **Running time:** the execution time of each software component, derived by instrumenting our prototype system.

We conduct the experiments with the abovementioned conferencing videos in our ultra-low-bitrate system (denoted as Our System in figures). Our system adaptively sends encoded images and landmarks according to the expected video quality. To our best knowledge, our system is the first complete system of its own kind, e.g., although Zeng et al. [23] also use facial models, they do not propose rate control mechanism, and thus their work cannot serve as the baseline system. For comparisons, we also run the same experiments with the x265 [22] codec, which is the state-of-the-art image-based codec (denoted as Image-based in figures). For each conferencing video, we vary the bitrate at $\{25, 30, 40, 50, 60, 100\}$ Kbps, and the base image size at $\{20, 40, 80\}$ KB. We let the bitrate be 40 Kbps and the image size be 80 KB if not otherwise specified. We run the experiments on a Linux workstation with an Intel i7 CPU at 3.6 GHz and 8 GB RAM. Our system works with different video quality metrics. Results from optimizing video quality in PSNR are shown by default; only sample results in SSIM are presented due to the space limitations.

4.2 Results

Our system outperforms the image-based system. Fig. 3 plots the PSNR values of a sample video under different bitrates. In both systems, higher bitrates lead to better conferencing video quality. However, our system constantly outperforms the image-based system. For example, our system achieves 30 dB[2] in PSNR at merely 25 Kbps, while the image-based system requires 60 Kbps (2.4 times). Next, we calculate the quality improvements in PSNR and SSIM of all conferencing videos and plot the mean improvements in Fig. 4. This figure shows that our system always outperforms the image-based system at all bitrates. The mean PSNR improvement of all conferencing videos is up to 7.5 dB, while the mean SSIM improvement is up to 0.08. For individual conferencing videos, the maximal PSNR improvement is 11.05 dB, and the maximal SSIM improvement is 0.15.

Our system adapts to network conditions and video content. We plot the average warping ratios at different bitrates in Fig. 5(a).

[1] To be realistic, we consider users who feel comfortable talking in diverse ways, e.g., some of them constantly move their heads, and others are rather static.

[2] It is considered as good quality [20].

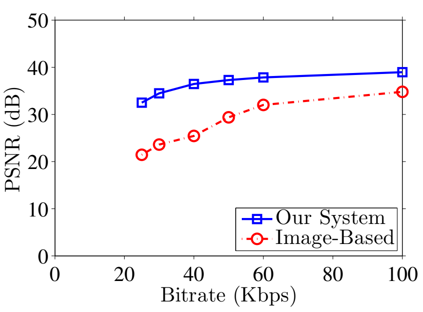
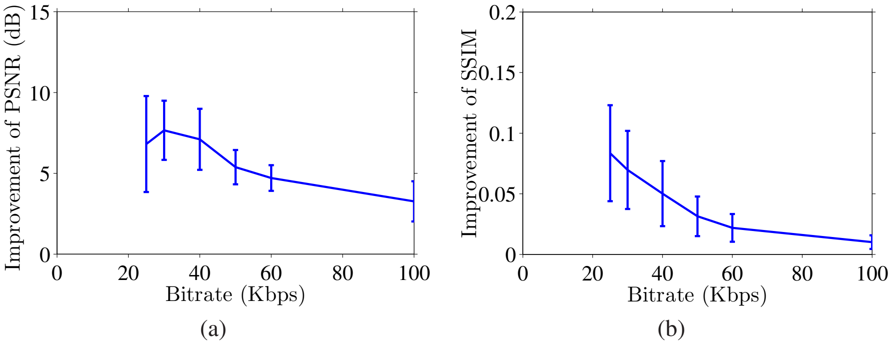


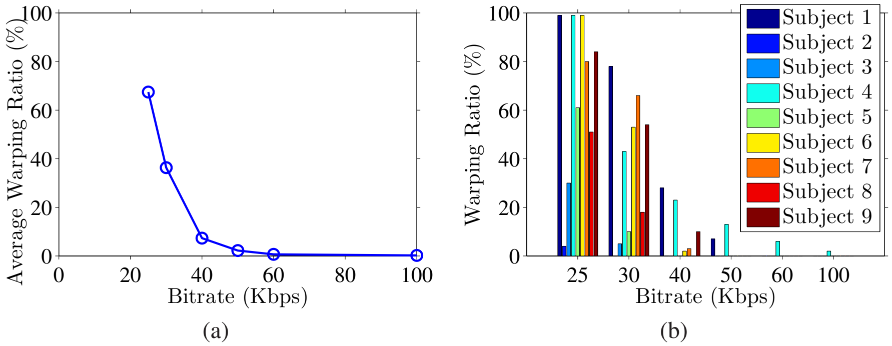
Figure 3: The video quality of a sample video under different bitrates.



(a)

(b)

Figure 4: Our system outperforms the image-based system in terms of conferencing video quality, in: (a) PSNR and (b) SSIM.



(a)

(b)

Figure 5: Our system automatically adapts to diverse: (a) network conditions and (b) video content.

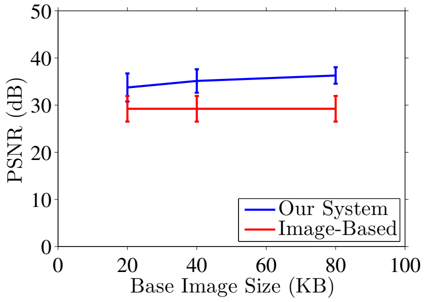


Figure 6: Better base image quality leads to higher conferencing video quality. Sample results in PSNR are shown.

In this figure, the average warping ratio increases as the total bitrate decreases. This shows the effectiveness of our system under different network conditions, because warped frames consume smaller bit budgets. Fig. 5(b) plots the ratio of the warped frames of all conferencing videos under different bitrates. This figure shows that our system results in diverse warping ratio with different video content. A closer look indicates that our system selects more warped frames when sending more active conferencing videos; and it selects fewer warped frames when sending more static ones. That is, our system is effective under different video content.

Implications of base image quality. We plot the average PSNR values across all conferencing videos with 95% confidence intervals in Fig. 6. We observe that, compared to the image-based system, our system achieves higher average video quality with all considered base image sizes, and the gap becomes slightly smaller if we reduce the base image size from 80 to 20 KB. More precisely, among all conferencing videos, our system outperforms the image-based system in 100%, 91%, and 83% of the conferencing videos, with the base image size of 80, 40, and 20 KB.

Running time. We measure and report the per-component running time. On average, the running times of the landmark detection, landmark encoding/decoding, image warping, image encoding/decoding, and the video quality assessment are 38, <1, 28, 13, and 5 ms, respectively. We note that some components are not invoked for every single video frame and some components may be pipelined. Our analysis reveals that the current (unoptimized) prototype system can achieves up to 12 fps. Several optimization techniques can be applied to further increase frame rate, e.g., H.265 codec chips may become commodity soon, which run much faster than the x265 software used in our experiment.

5. CONCLUSION AND FUTURE WORK

In this paper, we designed, implemented, optimized, and evaluated an ultra-low-bitrate video conferencing system that extracts facial landmarks, compresses and transmits facial landmarks and images, and warps the untransmitted images. Various optimization techniques were proposed to maximize the video quality. Using real conferencing videos, we conducted experiments to quantify the performance and limitations of our proposed system. The results show that our system: (i) outperforms the state-of-the-art x265 by up to 11.05 dB in PSNR, (ii) is able to adapt to different video content and network conditions, and (iii) runs in real-time at 12 fps. We believe the lessons learned when developing the proposed system will stimulate future research in this research area.

The presented work can be extended in several directions. First, multiple base images, e.g., with different facial expressions, may be sent and cached, so that the frame type selector may choose the base image that produces the highest warped frame quality. Second, a compression algorithm specifically designed for landmarks can be developed. Currently, landmarks are encoded using generic compression algorithm, leading to 21 Kbps bitrate on average. A customized compression algorithm that takes the landmark structure into considerations or even drops some less-critical landmarks may achieve lower landmark bitrate. Last, we plan to further speed up the individual components, especially the warper, using multi-threading, GPU, and techniques proposed in the literature [9, 15].

6. REFERENCES

[1] 7-zip official site, 2015. http://www.7-zip.org.

[2] N. Allen, B. Naidoo, and S. McDonald. Model-based compression for low-bitrate comms: A statistical approach to facial video encoding. In *Proceedings of Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*, Sep. 2006.

[3] A. Ansari and A. Mohamed. 3D face modeling using two views and a generic face model with application to 3D face recognition. In *Proceedings of IEEE Advanced Video and Signal Based Surveillance (AVSS)*, Jul. 2003.

[4] T. Baltruvsaitis, P. Robinson, and L. Morency. 3D constrained local model for rigid and non-rigid facial tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2012.

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, Aug. 1999.

[6] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[7] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):681–685, Jun. 2001.

[8] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *Proceedings of the British Machine Vision Conference (BMVC)*, Sep. 2006.

[9] V. Fuetterling, C. Lojewski, and F. Pfreundt. High-performance delaunay triangulation for many-core computers. In *Proceedings of the ACM SIGGRAPH Symposium on High Performance Graphics (HPG)*, Aug. 2014.

[10] I. Koufakis and B. Buxton. Very low bit rate face video compression using linear combination of 2D face views and principal components analysis. *Image and Vision Computing*, 17(14):1031–1051, Jan. 1999.

[11] D. Lee and B. Schachter. Two algorithms for constructing a Delaunay triangulation. *International Journal of Computer and Information Sciences*, 9(3):219–242, Feb. 1980.

[12] J. Ohm, G. Sullivan, H. Schwarz, T. Tan, and T. Wiegand. Comparison of the coding efficiency of video coding standards including high efficiency video coding HEVC. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1669–1684, Dec. 2012.

[13] Plan network requirements for Skype for business 2015, Sep 2015. https://technet.microsoft.com/en-us/library/gg425841.aspx.

[14] A. Puri and A. Eleftheriadis. MPEG-4: An object-based multimedia coding standard supporting mobile applications. *Mobile Networks and Applications*, 3(1):5–32, Jun. 1998.

[15] M. Qi, T. Cao, and T. Tan. Computing 2D constrained delaunay triangulation using the GPU. *IEEE Transactions on Visualization and Computer Graphics*, 19(5):736–748, May. 2013.

[16] X. Qi, Q. Yang, D. Nguyen, G. Zhou, and G. Peng. LBVC: towards low-bandwidth video chat on smartphones. In *Proceedings of ACM Multimedia System Conference (MMSys)*, Mar. 2015.

[17] D. Rathod, A. Vinay, S. Shylaja, and S. Natarajan. Facial landmark localization - a literature survey. *International Journal of Current Engineering and Technology*, 4(3):1901–1907, Jun. 2014.

[18] M. Suk and B. Prabhakaran. Real-time facial expression recognition on smartphones. In *Proceedings of the IEEE Applications of Computer Vision (WACV)*, Jan. 2015.

[19] Video conferencing market to expand at 9.3% CAGR to 2020 thanks to increasing usage in healthcare and defense, Jul. 2015. http://www.transparencymarketresearch.com/pressrelease/video-conferencing-market.htm.

[20] Y. Wang, J. Ostermann, and Y. Zhang. *Video Processing and Communications*. Prentice Hall, 2001.

[21] Z. Wang, L. Lu, and A. Bovik. Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication*, 19(2):121–132, February 2004.

[22] x265 HEVC Encoder official site. http://x265.org.

[23] W. Zeng, M. Yang, and Z. Cui. Ultra-low bit rate facial coding hybrid model based on saliency detection. *Journal of Image and Graphics*, 3(1):25–29, Jun. 2015.