ENABLING LIVE PRESENCE: DYNAMIC VIDEO COMPRESSION FOR THE TELEMATIC ARTS

Benjamin D. Smith
University of Illinois at Urbana-Champaign
National Center for Supercomputing Applications

ABSTRACT

Telematic performance, connecting performing artists in different physical locations in a single unified ensemble, places extreme demands on the supporting media. High audio and video quality plays a fundamental role in enabling inter-artist communication and collaboration. However, currently available video solutions are either inadequate to the task or pose extreme technical requirements. A new solution is presented, *vipr* (**v**ideo-**i**mage **pr**otocol), which exposes a number of popular, robust video compression methods for real-time use in Jitter and Max. This new software has successfully enabled several inter-continental performances and presents exciting potentials for creative, telematic artists, musicians, and dancers.

1. INTRODUCTION

Distributed, network performance, or *telematic performanc*e, is an exciting, growing area of aesthetic exploration. Musicians and dancers are engaging with remotely located counterparts on an increasing basis, testing the possibilities of live performance despite their physical separation. Distances both small and large are being overcome through the use of high-bandwidth networks, even enabling performances between ensembles distributed around the globe [2].

All live performance and inter-performer interactions require a sense of connection and presence in order to frame both verbal and artistic communication. Clear access to the movements, actions and sounds created by their remote partners is a prerequisite to coordinating expressive direction. Many media may be employed in facilitating this sense of presence, however video typically takes a pivotal role by relaying real-time footage between all of the remotely located participants. The quality of the video connection plays a significant role in creating this sense of presence and enabling (or denying) interactions between the performers [8].

While adequate audio technology has been available for some time [1] no satisfactory video software solution has previously been available. Typical performances employ easy-to-use programs such as Skype or iChat which provide poor performance quality, exhibiting large latencies, low color accuracy, frequent compression artifacts, low frame rates, and occasional connection loss. Other solutions use expensive, custom designed systems that require extensive expertise to operate. The setup of any system is further complicated by bandwidth requirements with drastically different capabilities depending on the available infrastructure.

Vipr for Max is a new, easy to use, yet powerful tool to enable live telematics performances, functioning as a dynamic image compression extension for Jitter, and is freely available for non-commercial use. This software provides a simple, robust interface to a variety of popular video codecs integrating seamlessly within the Max development environment. *vipr* uniquely provides a dynamic interface to the codec configuration, allowing users to tweak the bit rates, frame sizes, and codec selection in real-time, enabling ready identification of optimal settings as well as new areas for aesthetic expression through dynamic video deconstructing and editing.

2. MOTIVATION

The first musical collaborations over network connections were seen in the 1980s [3], employing satellite connections to bring artists across the United States into communication. Around this time the *League of Automatic Music Composers,* and its offshoot *The Hub,* formed with the express intent of exploring the potentials of network-based music and art [7]. Since the turn of the twenty-first century, with widespread institutional access to high-speed networks, artists have begun exploring telematic performance with much greater frequency. This has resulted in a plethora of examples [2, 3, 4, 5, 10], bridging all the performing arts as they intersect or extend into distributed, networked spaces. Musical works set in the networked domain come from one of two different approaches: the first focusing on the computers and network topography, seeking to employ them as instruments for artistic creation [7], and the second focusing on the communicative aspect of networks and their ability to bring people together across large physical (and

temporal) distances [10]. Many cases encompass both aspects, such as the work of Weinberg [11] and The Hub, however, the distinction between approaching computers *as* instruments versus enabling communication is significant.

Practitioners of the later have claimed the term *telematic* to denote works focusing on distributed, live performance that largely mimic conventional, western concert hall performance practices. These events typically involve ensembles comprising performing artists in two or more physical places connected by high-bandwidth networks relaying real-time audio and video. Thus the performers are able to engage with one another in collaborative performance, ideally transparently enabling co-present ensemble interactions. The technology to facilitate these performances is derived from computer-supported collaborative work systems and has the appearance of typical video conferencing setups.

However, fostering a sense of presence for the musicians requires a high degree of fidelity that extends beyond common telecommunications systems. *Presence* is the "the perceptual illusion of non-mediation" [8] and is required in order for a telematic performance to be successful. Further, Lombard and Ditton [8] identify a series of characteristics that are desirable for encouraging a sense of presence for the participants (who may be both performers or observers). These characteristics are: image quality, image size and viewing distance, motion and color, perceived dimensionality, and camera techniques.

Perceived image quality is reliant on many elements, including resolution, brightness, contrast, sharpness, color, and the absence of noise or artifacts. Higher resolution images tend to invoke greater presence, as does more photo-realistic images (which is a combination of accurate sharpness, color fidelity, contrast, etc.). Artifacts, typically resulting from image compression techniques, have also been found to decrease a sense of presence by drawing attention to the mediation of the experience.

Image size has similarly been shown to directly impact a viewers sense of presence, where images filling a larger field of view typically increase presence. Considering this problem in terms of field of view recognises that small images seem up close (such as in virtual reality goggles) can be equivalent to large images seen at a great distance (cinema screens, for example). In combination with the desire for higher image quality this typically equates to higher resolution images which increases the capability to display large, sharp, high contrast content.

Communication between performing artists requires conveying a sense of physical motion (i.e. halting sequences of still images do not suffice) and human visual perception studies [9] indicate that a sense of motion is most effectively created at video frame rates in excess of 50 frames-per-second (FPS), or 20 milliseconds per frame. The perception of motion is also convolved with image size and field of view to create the illusion of continuity.

For performing artists the amount of delay created by each leg of the network connection further impacts the sensation of presence. Shorter delays are preferred, although longer delays can be used successfully in certain situations [2, 10]. These delays, termed *latency*, also have far reaching artistic ramifications.

Consideration of the sensation of dimensionality (perceiving a three dimensional space in a two dimensional image) and the use of appropriate camera techniques (such as framing and shot length) are important to creating a sense of presence, however these operate independent of the video transmission system and thus are not treated further here.

Existing video solutions present a myriad of problems for the telematic performer attempting to create a reliable sensation of presence. These revolve partly around tradeoffs between bandwidth availability and computing resources, but also involve the amount of technical expertise required for operation. The easiest and most commonly employed systems (such as Skype) provide a readily accessible solution for the non-technically expert musician, however the image quality, size, and frame rate are all inferior (providing highly compressed, 640x480 video at under 25 FPS).

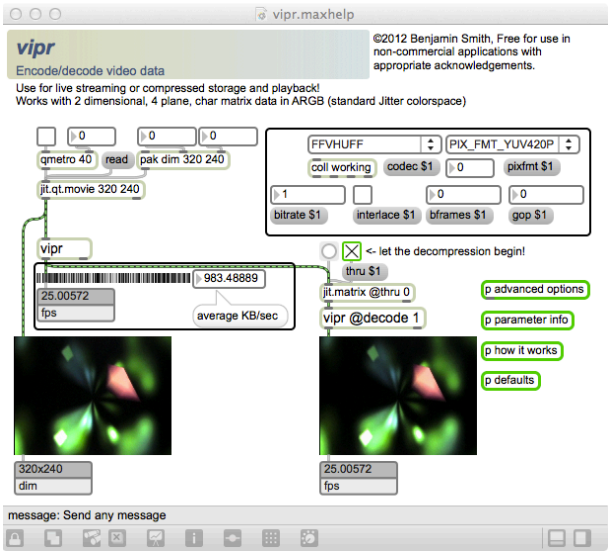

Figure 1. Screen capture of *vipr* in Max 6.

Technically advanced systems (such as that employed by [5], Access Grid, and SAGE) require dedicated technicians as well as expensive hardware

configurations. These solutions typically provide high resolution, high quality images (up to HD standard quality), but rely on access to very high-bandwidth institutional grade networks (internet2 and beyond). Mid-grade solutions (such as ConferenceXP, employed by [4]) still require a high degree of expertise, are operating system dependent, and provide poor hardware support.

3. VIPR

Vipr (fig. 1) was created to address the lack of flexible, accessible video transmission systems by providing access to a selection of popular compression/decompression (or *codec*) techniques within the Max/Jitter environment. This enables strong hardware support and independence, by leveraging Quicktime and DirectX, and takes advantage of the built in networking capabilities of Max. *Vipr* uses the open-source library *ffmpeg*, a very robust and highly successful video codec package that is continually incorporating new techniques and refinements, ensuring access to the best available compression models.

Within Max *vipr* operates as an independent object, compressing or decompressing individual matrices (i.e. images or frames of video) as they are sent to the object. The input matrices can be of any dimensions, in 32-bit color format. *Vipr* outputs a 1-dimensional matrix of image data that can then be stored or transmitted across a network to another instance of Max. When presented with an already compressed image matrix the external will decompress the image, outputting the original image. Processing takes place in real-time, yielding compression times on the order of milliseconds (depending on hardware and video image characteristics). Decompression is typically very fast (also on the order of milliseconds), and more extensive and precise metrics will be forthcoming. The resulting compression factor, or how many bits are saved by compressing the image, can be anywhere from 30% for lossless codecs (i.e. 70% of the original data space is saved) to 1% or less for lossy codecs (such as mpeg4). Due to the lightweight nature of *vipr* it is easily possible to setup multi-cast situations where each client in a network broadcasts its stream to multiple receivers.

Unique to this implementation, as compared to any other available video streaming solution, *vipr* exposes all of the parameters of the codecs for real-time control within Max. This not only enables rapid stream optimization capabilities but also presents the potential for the artistic use of the compressors by intentionally pushing the system into unusual states (such as creating artifacts in the image by dropping or repeating frames of the video while changing the bit rate).

The lossless codecs made available by *vipr* (such as *huffyuv*, *ffv1*, and *ffvhuff*) present another unique option for the telematic artist. These codecs provide fully accurate image reproduction on the receiving end without any compression effects in the image. This is a significant advantage when high image quality is desired (in order to facilitate the sense of presence), and their real-time implementation is currently unique to *vipr*. While the processing load is slightly higher (requiring around 20 milliseconds for a 720p image on typical modern Mac hardware) the data reduction (on the order of 8:1, yielding a stream of 25.6 megabits-per-second for a 720p stream) enables transmission over networks that cannot support raw HD video streams (which would typically require a 10 gigabit network or better).

3.1 Codec Parameters

The principle codec parameters exposed by *vipr* are now described. These are only set on the compressing side of the network, as the decompressing system can automatically detect the settings used for compression.

Bit-rate–This sets the target compression amount a lossy codec will attempt to achieve. The rate is not guaranteed, as most codecs (such as the mpeg series) use a variable compression scheme to take advantage of highly compressible sequences, and the final rate may be higher or lower than the target.

Group Of Pictures (or *gop*)–Lossy codecs typically employ a key framing technique in order to minimize compression artifacts in unreliable situations (where frames may be corrupted or dropped). The key frame is a single image with a fully descriptive encoding, allowing the original image to be constructed from the key frame alone. Key frames are typically sent as one out of every 10 frames or more, depending on the reliability of the network connection. The frames sent between the key frames are incremental in nature, requiring the previous image in order to recreate the source image. This typically results in greatly improved compression rates (as any constant pixels between two images in a sequence do not need to be retransmitted) yet is very susceptible to lost frames which will produce noticeable image quality degradation. The *gop* parameter sets the number of incremental frames to use between key frames.

B-Frames–While the incremental frames resulting from the *gop* setting only rely on the previous image state (and are termed *i-frames*), b-frames use a bi-directional model that compresses based on both the preceding and following images. This provides even greater compression rates. However, b-frames also require delaying the stream by the same number of frames, because compression cannot complete until the following image has been processed. While this is

perfectly acceptable for prerecorded video sequences it may be undesirable in real-time settings that attempt to minimize latency in order to further the sense of presence.

Pixel format–This dictates how many bits are used for each color component in the image during the internal compression process. Most codecs only operate with images that are in one of a few specific pixel formats, while Jitter uses a 32-bit format that is inefficient and not supported by many codecs. Thus *vipr* transparently converts the image internally for compression, returning it to a 32-bit format upon decompression. Typical internal formats use fewer than 32 bits, providing additional compression advantages, but shifting the color space slightly (which can be detectable in some situations). Codecs that support multiple pixel formats allow the user to change the internal format, which may result in minor image quality improvements.

Resample method–In order to perform the pixel format conversion a resampling method is used, which may be selected by the user. *vipr* implements a variety of methods (including *bilinear, fast bilinear, bicubic, point, area,* and *gaussian*), which have measurable performance impacts. Typically, *fast bilinear* provides optimal efficiency however other methods may be preferable depending on the content being converted.

4. DISCUSSION

Vipr provides customizations to enable the best recreation of presence possible for nearly any hardware configuration. Recently it has been used as a component in several inter-continental performances with great success [10] and is being employed as a key technology in a telematic opera production [4].

Finding the optimal parameter configuration is highly dependent on the desired results as well as the available computation resources (i.e. CPU time), network bandwidth, camera quality, and projection capabilities. At the moment general guidelines are not available, however this is a focus of ongoing research. Yet, the immediate response of the system allows even novice users to explore parameter settings in order to maximize quality within a given setup.

Adjusting for the preferred setup requires balancing image size, image quality, and frame rate with the resulting processing load and bandwidth allowances. Ideally a telematic performance situation would involve large-as-life or larger-than-life projections on a stage alongside the 'live' musicians, operating at upwards of 50 FPS with highly accurate color reproduction. However this relies on expensive equipment, high grade networks, and precludes many small or more highly distributed events.

Yet studies with live musicians have found that unnoticeably compressed images may not have a significant impact on the fostered sense of presence [10]. Without the appearance of noticeable compression artifacts the loss of sharpness and color smearing apparently has minimal implications for a musician's sense of connection through a tele-present system. Additionally, employing similar codecs opens up the possibility of using consumer grade networks for professional performances. While large projections warrant HD quality video streams and can benefit from high-bandwidth networks personal monitors can easily take advantage of 480p or smaller image sizes.

Perhaps the most valuable aspect for fostering a sense of presence between non-co-located performing artists is the perception of motion [10]. Thus frame rates operating at the theoretically optimal human processing rate are key (with new images every 20-30 milliseconds, or 33-50+ FPS) [9]. Previous solutions (barring a few highly expert systems) have been unable to provide this functionality, especially for lower frame sizes and higher compression rates. *vipr* makes no frame rate assumptions, being limited solely by the available hardware, and easily performs at over 100 FPS for smaller image sizes on contemporary PCs.

Informal evaluations have found that both frame rate (typically described as the fluidity of motion) and color fidelity (i.e. the vivid nature of the images) are the primary aspects of *vipr* setups that new observers comment on.

While latency is typically considered the primary problem facing telematic performing artists, *vipr* does little to alleviate this challenge. Latency has many components creating the overall effect, including image capture and digitization, compression, transmission, reassembly and decompression, and projection time. The primary factor in transmission time is the physical distances involved, where a single packet always takes a minimum amount of time, and even the speed of light sets hard limits on potential delays. However, *vipr* provides a profitable tradeoff between increased compression times (on the order of milliseconds) and reduced transmission times. While an individual packet still takes the same time to deliver, the amount of time required for the transmission of the whole frame is reduced proportional to the compression amount (anywhere from 1% to 30%), which can lead to significant increases in performance. Noisy situations with high packet loss rates benefit even more, as far fewer packets are required to send the compressed images and thus far fewer are lost and require retransmission.

The network transmission of *vipr* images relies on the Max built-in network objects (*jit.send* and *jit.receive*). These use the TCP/IP protocol which has both advantages and limitations for video transmission. The primary benefit is the guaranteed in-order delivery of frames, ensuring complete reception of the video on the destination machine. However, this requires the retransmission of dropped packets which can cause increased latency (as the stream must be buffered and wait while the missing packet is requested and retrieved). The alternative protocol, UDP, does not have implicit quality control and thus does not retransmit lost packets, but packets may arrive out of order or be lost entirely resulting in significantly degraded video quality. At the moment TCP provides the only reliable solution, especially for the non-expert artist.

5. CONCLUSION

Telematic performing artists rely on a sense of presence in order to create the desired collaborative works and this sense of presence can be facilitated by live video connections. A number of characteristics directly impact the formation of presence: image quality, image size, color fidelity, and motion are all key elements of the video system. However, typical solutions rarely provide enough flexibility for the artist, requiring technically complex custom solutions to create a desirable telematic setup.

Vipr for Max provides a ready solution to this problem and has been successfully employed by musicians and dancers in several trans-atlantic and trans-pacific concerts. This freely available compression object for Max and Jitter opens up new doors for telematic artists, enabling high quality, highly flexible video transmission across smaller network paths. In a uniquely interactive paradigm, *vipr* allows the user to experiment with codec settings in real time, in order to both locate the optimal configuration and to expose new aesthetic styles.

6. ACKNOWLEDGEMENTS

The author would like to thank Dr. Guy E. Garnett, the Illinois-Japan Performing Arts Network, eDream, and the National Center for Supercomputing Applications for supporting this work.

7. REFERENCES

[1] Cáceres, J.P., and C. Chafe. “JackTrip: Under the hood of an engine for network audio.” *Journal of New Music Research* 39, 3: (2010) 183–187.

[2] Cáceres, J.P., R. Hamilton, D. Iyer, C. Chafe, and G. Wang. “To the edge with china: Explorations in network performance.” In *ARTECH 2008: Proceedings of the 4th International Conference on Digital Arts*. 2008, 61–66.

[3] Cooperstock, Jeremy. “History of Spatially Distributed Performance”, 2011, Accessed June 8th, 2011. http://www.cim.mcgill.ca/sre/projects/rtnm/history.html.

[4] Deal, S. “Auksalaq, a Telematic Opera”, In *Proceedings of the International Computer Music Association Conference 2011*.

[5] Dresser, M. “Telematics.”, 2008, Accessed Dec 22, 2010. http://www.allaboutjazz.com/php/article.php?id=30198.

[6] van Eijk, R.L.J. *Audio–visual synchrony perception*. Unpublished doctoral dissertation, Technische Universiteit Eindhoven, 2008.

[7] Gresham-Lancaster, S. “The aesthetics and history of the hub: The effects of changing technology on network computer music.” *Leonardo Music Journal* 8: (1998) 39–44.

[8] Lombard, M., and T. Ditton. “At the heart of it all: The concept of presence.” *Journal of Computer-Mediated Communication* 3, 2 (1997).

[9] Pastor, M.A., and J. Artieda. *Time, internal clocks, and movement*. North Holland, 1996.

[10] Smith, B. *Telematic Composition*. D.M.A. thesis, University of Illinois at Urbana-Champaign, 2011.

[11] Weinberg, G. “The aesthetics, history, and future challenges of interconnected music networks.” In *Proceedings of the International Computer Music Association Conference*. 2002, 349–356.