POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

Real-time and high-quality video compression for telesurgery

TESI MAGISTRALE IN BIOMEDICAL ENGINEERING – INGEGNERIA BIOMEDICA

AUTHOR: MARTINA GOLINI

ADVISOR: PROF. ELENA DE MOMI

COADVISOR: IURI FROSIO, ALDO MARZULLO

ACADEMIC YEAR: 2021-2022

1. Introduction

Nowadays the advances in telecommunication technologies and video compression systems have led to the development of new way to deliver high-quality cares and new teaching strategies. Telementoring, telemedicine and telesurgery are the new frontiers of medicine, which enables for diagnosis and surgeries in different contexts, e.g., in disaster-affected and distant rural areas. In this scenario, an experienced surgeon/physician can offer a real time support to less-skilled colleagues; moreover, it can be performed a surgery from a remote position, and it is given the opportunity for students to watch and learn. The new realities of medicine address new challenges, since a big amount of data need to be stored and transmitted. Even though the leading standards for video compression, i.e., H.264/AVC and H.265/HEVC, are highly optimized and performing, the strict constrains required by these applications have brought to explore brand-new solutions either for the optimization of the traditional methods and to develop alternative strategies. In this perspective, Deep learning-based techniques have been exploited for the purpose since they can overcome the limitations of standard systems. Focusing on video transmission, it presents latency and bandwidth constrains to guarantee a real time application, together with the need to preserve the quality. In details, the threshold for real time application is set to **30ms** for encoding/decoding time. In this work, it is proposed a computational friendly, deep learning-based scheme to jointly satisfy the request of low-latency and bandwidth and high-quality compression.

More in detail, it is implemented an autoencoder for the coding of the residual, i.e., the difference between the original frame and the compressed one, obtained by employing the H.264/AVC codec, since it has been demonstrated the leading standard in the surgical domain. The output of the neural network is eventually summed to the one of H.264/AVC for a better reconstruction of the images. The research focuses on Robotic Assisted Minimally Invasive Surgery (RAMIS), which is spreading among various surgical area [1] and requires for high-quality and low latency, to guarantee the stability of the system employed. For the aim of the work, it is utilized **720p** - thus HD resolution- Robotic Assisted Radical Prostatectomy (RARP) videos obtained from the Da Vinci robot, since RARP represents one of the most performed RAMIS operations. The scheme proposed shows to overcome H.264/AVC performances in a low bitrate scenario, allowing for high quality and real time applications, especially in all those contexts featuring a poor Internet connection.

2. Related works

Among the many traditional video codecs, H.264/AVC and H.265/HEVC are the most adopted and diffused. Both these codecs are based on the hybrid prediction/transform coding method, first proposed in 1979 [2]. Despite H.265/HEVC overcomes its predecessor in terms of performances, it is less hardware-friendly, thus H.264/AVC remains widely employed for many applications, including the surgical ones. Since both codecs utilize a block-based scheme, they are interested by block artifact and quality degradation due to the quantization process. For these reasons, DL-based codecs have started to be explored as a promising alternative. Researchers have exploited DL techniques either to design brand-new schemes and to optimize one of the main modules of the traditional codecs, i.e., intra-prediction, inter-prediction, quantization, entropy coding and loop filtering [2]. Lu et. al [3] have developed a DL scheme which substitutes each module of H.264/AVC with a Convolutional Neural Network, achieving the state-of-art results at the time of publication. While this solution modifies the entire traditional pipeline, many others have focused only on one phase of the standard framework. Li et al.[4] have proposed a five layers CNN-based block up-sampling scheme to improve the intra prediction module. This scheme achieved an important reduction in terms of required bandwidth, but it requires a significantly higher encoding/decoding time than H.265/HEVC. Feng et. al [5] have built an enhancement module which operates before a super-resolution network to deal with sampling and compression artifacts separately. In[6] it is proposed a binary autoencoder to generate a binary code which is transmitted together with the frame data. On the decoder side, it is performed a residual correction of the image compressed by H.264/AVC. In the surgical domain the usage of DL-based techniques for video compression and transmission is poor documented. It has been demonstrated that the detection of clinically relevant spatio-temporal information can be exploited to save compression time. Therefore, CNN can be used for the recognition of the Regions Of Interest (ROI) [7].

The Summary is organized as follow: section 3 presents the solution proposed for real time and high-quality video compression; section 4 contains the results and the discussion; section 5 is left for the conclusions.

3. Binary Residual Neural Network for RARP video compression

In the following section the structure of the scheme proposed, as well as the dataset

employed and the training process is described.

The binary output is sent to the decoder, which performs the up-sampling. The decoder

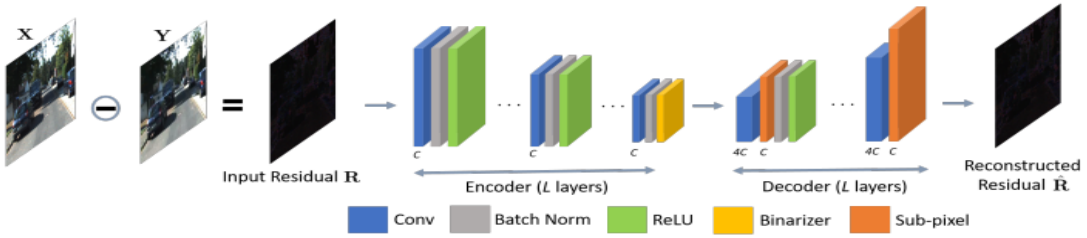

Figure 3.1.1 The pipeline of the proposed method

3.1 The network architecture

The scheme proposed (Figure 3.1.1) has been designed by *Tsai et. al* [6]. The input **X** has been compressed by H.264/AVC, to generate the output **Y**. The difference between the original frame and the coded one, i.e., the residual **R**, is compressed by a binary autoencoder. The output of the autoencoder **R'** is eventually summed to **Y**, producing the final reconstructed frame. The autoencoder features three functions: an encoder **E**, a binarizer **B** and a decoder **D**. The first one extracts compact features representations, which are eventually sent to the binarizer. The encoder is composed by L 2D-convolutional layers, with equal number of channels C, characterized by a stride of 2, which performs the down-sampling. The binarizer maps each element e_i received in the interval [-1,1] and discretize it to {−1, 1}, producing a binary output. To the purpose, it is employed two different functions, i.e., the activation function σ and the discretization function b:

$$B(e_i) = b\big(\sigma(e_i)\big) \quad (3.1)$$

For its superior performances, the *hardthan* is employed as binarization function:

$$b(z) = \begin{cases} 1, & if\ z > 1 \\ z, & if -1 \le z \le 1 \\ -1, & if\ z < 1 \end{cases} \quad (3.2)$$

With $z = \sigma(e_i)$.

consists of L 2D-convolutional layers, each one of them followed by a SubPixel, layer featuring an upscaling factor of 2; the convolution process and subpixeling are jointly employed for up-sampling. In this case, the number of channels used in the first two convolutional layers is equal to 4xC, due to the presence of the SubPixel layer. For the same reason a stride of 1 is employed. The last layer of the decoder, instead, presents 12 output channels C Both the encoder and the decoder present ReLu as activation function. An additional operation, i.e., batch-normalization with a momentum of 0.999 is also included to facilitate the learning process. The kernel size is set to 2 for the convolution operations during encoding, while is set to 1 for each convolutional layer of the decoder. The size of the binary map strictly depends upon the width W and the height H of the input image as well as on the number of channels C and the layers L which characterized the neural network. Specifically, the size is given by:

$$S = \frac{C \times W \times H}{2^{2L}} \quad (3.3)$$

Intuitively, a deeper neural network corresponds to a smaller binary map, thus the compression task would be easier, while the training would be harder. Therefore, C and L needs to be carefully chosen to guarantee a good trade-off between these two processes. Based upon the studies developed by *Tsai et.al* [6], the number of channels C is set to 32, while the number of layers is set to 3.

3.1. The dataset

For the aim of the work, five high-quality videos with the endoscopic view captured during RARP (**1280 x 720**) are downloaded from YouTube. The video duration ranges from 72 up to 100 minutes. The first video (Video A) is used for *testing*, while the other three (Video B, C, D, E) are used for *training*. To highlight the different phases of the procedure from each video, ten 40 seconds clips are selected. They include different anatomical sections, surgery instruments, levels of illumination and degrees of action performed in the surgery field. The clips extracted are eventually compressed and decompressed by using the H.264/AVC implementation provided by FFmpeg, with a particular focus on bandwidth and latency, both dependent on the bitrate and the preset selected. The FFmpeg preset represents the coding speed value, thus returns a certain compression ratio / frame quality / compression time. For the aim of the research, Ultrafast, Medium and Slow presets are employed. Compressing at different bitrate allows investigating the codec performance as a function of the transmission bandwidth. In this work the first evaluation is conducted employing three bitrate values, i.e., 1,2,5Mb. Thus, 9 bitrate/preset pairs are evaluated. A further configuration, i.e., 10Mb-Ultrafast, is eventually investigated based on the results obtained from the first analysis. From each videoclip one frame every ten is extracted, thus each testing dataset -one for each bitrate/preset pair is composed by 1216 frames, while each training/validation dataset is formed by 4802 frames (70% training set/30% validation set). Both original and compressed frames are employed for the residual computation, which is coded by the autoencoder.

3.2. Training the residual autoencoder

The autoencoder is implemented in Python, using the PyTorch library. The training hyperparameters, i.e., learning rate η, batch size and number of epochs, are set respectively to 0.01, 5 and 50. Moreover, the learning rate is reduced by a factor of 0.5 every 5 epochs; in fact, decreasing the learning rate during training can lead to improved accuracy and reduced overfitting of the model. It is used Mean Square Error (MSE) as loss function, while Adam is employed as optimizer. The training is performed by using the NVIDIA GeForce GTX 850M GP.

4. Performance evaluation

To assess the performances in terms of quality the Peak-To-Noise-Ratio (PSNR) and the Structural Similarity (SSIM) [8] are computed for both images obtained by H.264/AVC and by the scheme implemented, employing the original frames as ground truth. The metrics are reported in function of bit-per-pixel (BPP):

$$BPP = \frac{80000 \times Bitrate\,(Kbs)}{H \times W \times fps} \quad (4.1)$$

Moreover, it is measured the encoding/decoding time for both traditional and DL-based techniques to evaluate latency.

4.1. Results

In terms of **quality** the scheme proposed outperforms the traditional standard H.264 in a low bitrate scenario.

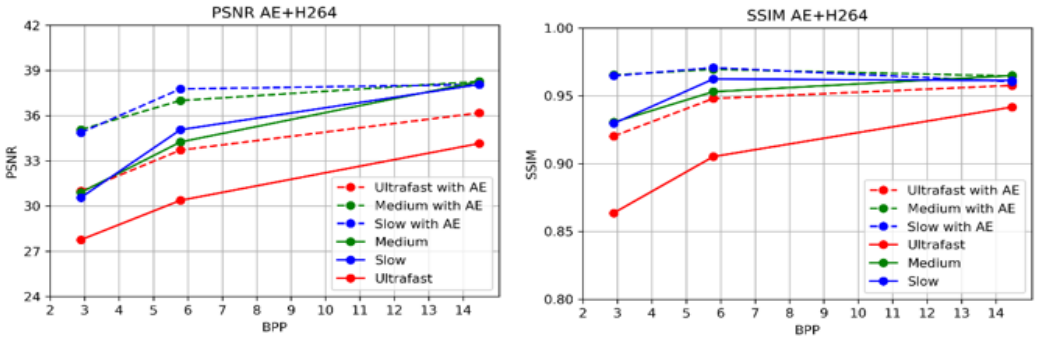

Figure 4.2 The plots show a comparison between the codec standard H.264 and the proposed scheme in terms of quality. Both PSNR (right) and SSIM (left) are expressed in function of the bitrate.

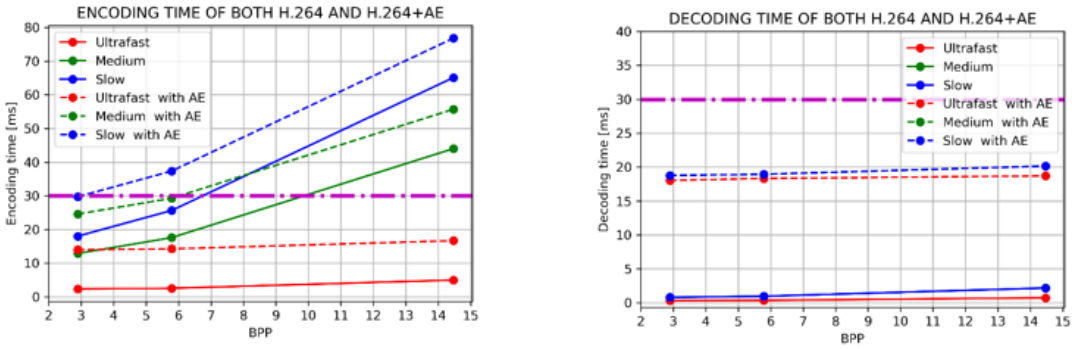

Figure 4.1 The plots report a comparison between the performances in terms of encoding/decoding time of H.264 and of the scheme implemented, for each configuration. The purple line represents the threshold for real time applications.

As (Figure 4.1) indicates, the mean PSNR value of each bitrate/preset pair is comprised between 30 dB and 38 dB for both H.264/AVC and the method proposed, except the one associated to the 1Mb-Ultrafast configuration for H.264/AVC. Since typical PSNR values for 8-bit data range from 30 dB to 50 dB, it can be stated that the quality of the reconstruction performed by the scheme proposed is on average good (PSNR values > 30 dB) and better than the one achieved with H.264/AVC, except for the 5Mb-Medium/5Mb-Slow/10Mb-Ultrafast pair, which is demonstrated to be equal. SSIM indicates the perceived quality of digital images and video. Its values range from 0 to 1, where 1 indicates the perfect structural similarity. Figure 4.1 demonstrates that the perceived quality is on average meaningly better for the frames reconstructed by the DL-based scheme for a bitrate equal to 1 Mb and for the 2Mb-Ultrafast configuration, while is almost unnoticeable for higher bitrate, i.e., 2 Mb and 5 Mb. Moreover, the SSIM values related to 5Mb-Medium/5Mb-Slow pair indicates a slightly higher perceived quality for images compressed by the traditional codec. The perceived quality is high for frames reconstructed both by H.264/AVC and the scheme proposed; therefore, it can be visually noticed a difference in the images compressed by using the ultrafast preset. A further analysis to state the reliability of the results is conducted by employing the Mann-Whitney U test. It is demonstrated that there is a difference between the traditional and the proposed method for almost the entire bitrate/preset set, with an exception for the 5Mb-Medium/5Mb-Slow/10Mb-Ultrafast pair, for which the p value is respectively 0.86, 0.90 and 0.79 and the null hypothesis is not rejected.

Time presents a superior limit, as low latency is requested to guarantee real time applications. More in detail, the threshold is set to **33,3ms** (30 Hz) per frame, for both encoding and decoding time. In (Figure 4.2) the threshold is indicated by the purple line.
Traditionally, the time requested for the encoding process is significantly higher than the one addressed to the decoding one. However, the method implemented shows opposite results, since the decoder is more computational demanding. As for the 10Mb-Ultrafast pair, the encoding/decoding time remains lower than 30ms.

4.2. Discussion

Analyzing the mean values obtained both for PSNR and SSIM for each clip, which present a common trend, it is observed that significantly lower values are always obtained for clip 5 and clip 7 thus it is conducted a frame-by-frame evaluation. it can be stated that the deep learning-based scheme performs lower quality compression where significantly fast movements are present. To select the most suitable bitrate/preset pair, it needs to be considered the encoding/decoding time, since the latency of the video feedback highly limits telesurgery applications. It needs to be highlighted that the delay between the movement performed by the surgeon through the master console and its visualization on the video screen it is composed by the sum of latency due to the video codec and the one associated to the transmission signal which allows the motion. In literature is found 330ms to be the maximum value recommended for telesurgery, where the latency associated to the video codec it is 70ms (encoding + transmission + decoding) [9]. Based on this, it is selected 1Mb-Slow, 2Mb-Medium and 10Mb-Ultrafast as they represent the best quality-time trade-off for that bitrate value. The most suitable bitrate/preset pair results 1Mb-Slow for it achieve the highest perceived quality, while not overcoming the time threshold. The configurations chosen are able to transmit 30 frame per second, even if in some cases the sum between encoding and decoding time overcomes the threshold. The real time application remains possible assuming that both encoder and decoder are working at maximum 33,3ms each - without considering the transmission time - since encoder and decoder run on different devices. It is worth noticing that the scheme proposed is not developed for speed compression, while H.264/AVC is highly optimized for the purpose. Besides, it is used a low-performing GPU, thus the computation could be accelerated employing a better one. It is also to be considered that the dataset is downloaded from YouTube, thus videos have been previously compressed. It is clear that the implemented method may be widely optimized to achieve better performances both in terms of quality and speed.

5. Conclusions

This work presents a computational-friendly solution for surgical video compression which is capable to jointly enhance the compression quality and work under low-latency constrains in a low bitrate scenario. In other words, this scheme offers the possibility to obtain good compression quality of high-resolution videos in a low-bandwidth domain, which is useful in all those contexts that feature a non-fast internet connection, e.g., developing Countries and rural areas. The quality guaranteed is high, thus it allows for the detection of every detail in the surgical area in different situations, e.g., bleeding and smoking. Even though the reconstruction of really fast movement is more difficult, the quality perceived do not compromise the

result of the surgery. The solution proposed allows for remote surgery in which the distance between the surgeon and the patient could be of more than 14 000 km, since latency can remain considerably under 70ms. The method implemented can be widely modified to become a powerful tool for telemedicine, telementoring and remote surgery applications. In fact, further optimizations could make the network more performant, especially in terms of speed. Moreover, many surgical procedures exploit stereo-images to enable 3D perception. Even though the transmission is more complex, it could be leveraged the redundancy between the left and right images to develop highly performant solutions among different field other than the surgical domain, such as virtual reality and videogames. The progress in compression systems may lead to the spreading of tele-health, which can have a strong impact on the quality of life and may improve the learning process of medicine students.

6. Bibliography

[1] J. H. Palep, "Robotic assisted minimally invasive surgery," *Journal of Minimal Access Surgery*, vol. 5, no. 1, p. 1, Jan. 2009, doi: 10.4103/0972-9941.51313.

[2] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and Video Compression With Neural Networks: A Review," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1683–1698, 2020, doi: 10.1109/TCSVT.2019.2910119.

[3] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An End-to-end Deep Video Compression Framework", Accessed: Feb. 25, 2022. [Online]. Available: https://github.com/GuoLusjtu/DVC.

[4] Y. Li *et al.*, "Convolutional Neural Network-Based Block Up-Sampling for Intra Frame Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2316–2330, Sep. 2018, doi: 10.1109/TCSVT.2017.2727682.

[5] L. Feng, X. Zhang, X. Zhang, S. Wang, R. Wang, and S. Ma, "A Dual-Network Based Super-Resolution for Compressed High Definition Video," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11164 LNCS, pp. 600–610, Sep. 2018, doi: 10.1007/978-3-030-00776-8_55.

[6] Y.-H. Tsai, M.-Y. Liu, D. Sun, M.-H. Yang, and J. Kautz, "Learning Binary Residual Representations for Domain-specific Video Streaming", Accessed: Mar. 04, 2022. [Online]. Available: www.aaai.org

[7] S. Khire, S. Robertson, N. Jayant, E. A. Wood, M. E. Stachura, and T. Goksel, "Region-of-interest video coding for enabling surgical telementoring in low-bandwidth scenarios," *Proceedings - IEEE Military Communications Conference MILCOM*, 2012, doi: 10.1109/MILCOM.2012.6415792.

[8] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," *Proceedings - International Conference on Pattern Recognition*, pp. 2366–2369, 2010, doi: 10.1109/ICPR.2010.579.

[9] S. E. Butner and M. Ghodoussi, "Transforming a Surgical Robot for Human Telesurgery," *IEEE Transactions on Robotics and Automation*, vol. 19, no. 5, pp. 818–824, Oct. 2003, doi: 10.1109/TRA.2003.817214.