Frame Interpolation Using Generative Adversarial Networks

Mark Koren
mkoren@stanford.edu

Kunal Menda
kmenda@stanford.edu

Apoorva Sharma
apoorva@stanford.edu
Stanford University 496 Lomita Dr. Stanford, CA 94305

1. Introduction

Video services, especially streaming services, are some of the most recognizable brands in technology today. One of the hardest problems in this exciting field is that of frame interpolation.

Frame interpolation, for the purposes of this project, is the action of generating a frame for a video, given the immediate frames occurring sequentially before and after. This allows a video to have its frame rate enhanced, which is a process known as *upsampling*. In the general upsampling task, we cannot assume access to ground truth for the interpolated frames.

High fidelity upsampling can be applied to video *compression*, since could store only key frames of the video, and interpolate at playtime to fill in the missing parts. For compression tasks, the original high frame rate video exists, and thus the ground truth for the interpolated frames is available.

Inspired by the successes of end-to-end deep convolutional networks in outperforming conventional techniques for image classification, we propose an end-to-end neural architecture ("FINNiGAN") for the frame interpolation task. The input to this algorithm is a pair of sequential frames from a video. We use a convolutional neural network (CNN) architecture involving the generative adversarial network setup to generate the frame which would appear temporally between the input frames.

In the following sections, we first discuss related work, then outline our methods, introduce the dataset used for testing, and finally discuss results.

2. Related Work

There are several conventional image processing techniques for video frame interpolation. The simplest method for interpolating between two frames is *Linear Frame Interpolation*. In this technique, for each pixel location x in the interpolated frame, the value is linearly interpolated between the neighboring frames:

$$I_{1/2}(x) = \frac{1}{2}(I_0(x) + I_1(x))$$

As this is a pixel to pixel method, it fails to properly account for motion of objects across pixels. This creates an effect known as "ghosting" where objects that are in motion have multiple edges in the interpolated frame.

Current state of the art frame interpolation is done using an algorithm called Motion-Compensated Frame Interpolation (MCFI), which is currently used in many HDTVs []. MCFI techniques work in two parts: Motion Estimation (ME) and Motion Compensation (MC). ME often involves computing the 'velocity' of each pixel in the frame, i.e. how a given pixel's content shifts between the frames.[] At a high level, the MC step involves using these motion estimates to move each pixel halfway in the same direction. []. These steps perform significantly better than LFI, but suffer from their own artifacts, such as "tears" or misplaced blocks, resulting in qualitatively unsatisfactory results described as having a "soap-opera effect."

Recent work by Guo and Lu [] presents an improvement to MCFI called I-MCFI, and also gives a survey of other state-of-the-art frame interpolation techniques, such as Adaptive Vector Median Filtering (AVMF) and Motion Vector Smoothing (MVS). Our work will be compared against these algorithms as a baseline.

At its core, frame interpolation is a two image to single image translation task, which involves making sense of information from two images, and then generating a single image. Convolutional neural networks (CNN) can be applied in an encoder-decoder setup to learn implicit features in images. [**?**] Previous work [1] utilized a CNN architecture for frame interpolation task and achieved promising results. The primary issue with the pervious work was the blurriness and noisiness of the generated images.

Generative Adversarial Networks (GANs) [**?**], have been shown to be very good at realistic image generation [**?**]. Conditional-GAN [**?**] and pix2pix [**?**] adapt the

GAN framework to single-image-to-single-image translation, achieving good results in going from outlines and cartoons to photorealistic images. In this paper, we build on the work in [1] by incorporating a GAN architecture to improve the photorealism of the result.

3. Methods

Our method performs **F**rame **I**nterpolation with a Convolutional **N**eural **N**etwork as well as **i**ncorporating **G**enerative **A**dversarial **N**etworks for image refinement, hence the name "FINNiGAN."

4. Dataset

5. Results

6. Conclusion

References

[1] A. Sharma, K. Menda, and M. Koren. Convolutional neural networks for video frame interpolation.