Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation

Jose Caballero, Christian Ledig, Andy Aitken, Alejandro Acosta, Johannes Totz, Zehn Wang, Wenzhe Shi
{first name initial}{surname}@twitter.com

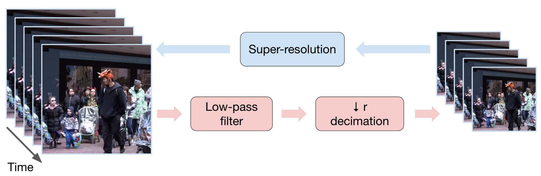
Twitter

#KeyIdea

We propose a real-time, accurate and temporally consistent super-resolution method for 1080p 30fps video.

#Introduction

Video super-resolution (SR) estimates a high-resolution video from its low-resolution version.



The problem is ill-posed and reconstruction usually exploits spatio-temporal redundancies. Previous approaches have been **inefficient** (sub real-time speeds) [3] or **naive** (treat frames independently) [1].

Contributions

An end-to-end trainable convolutional neural network for joint frame motion compensation and video SR, improving:

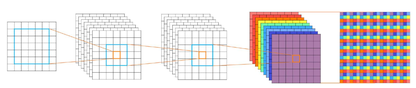
Efficiency Processing is done in LR space and mapped to HR space with sub-pixel convolution

Accuracy
- Spatio-temporal architecture exploits correlations in space and time
- Motion compensation further exposes temporal redundancies

#Background

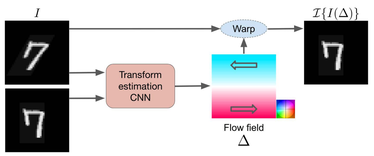
ESPCN [1]

Direct mapping of LR to HR images with sub-pixel convolution.



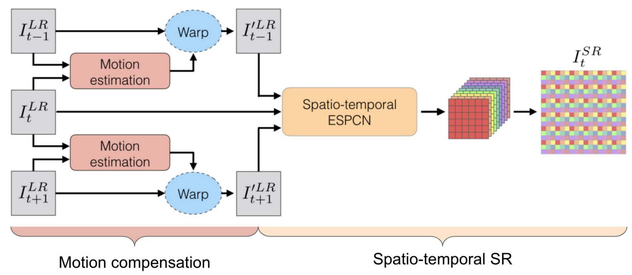
Spatial Transformers [2]

Learning image transformations.



#Method

A convolutional neural network (CNN) processes an odd number of consecutive frames to estimate the SR middle frame.



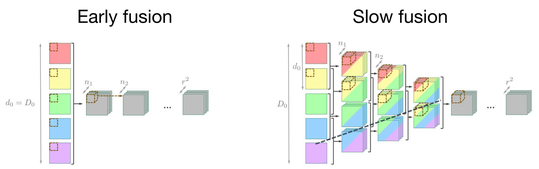
We train model parameters that jointly minimise the error of the HR frame reconstruction and the motion compensation from neighbouring frames.

$$(\theta^*, \theta^*_\Delta) = \arg\min_{\theta,\theta_\Delta} \|I_t^{HR} - f(I'^{LR}_{t-1:t+1}; \theta)\|_2^2 + \sum_{i=\pm 1} [\beta \|I'^{LR}_{t+i} - I_t^{LR}\|_2^2 + \lambda \mathcal{H}(\partial_{x,y} \Delta_{t+i})]$$

Spatio-temporal SR — Motion compensation

Spatio-temporal networks

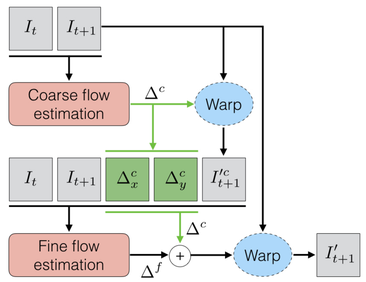
We study different approaches to process temporal information.

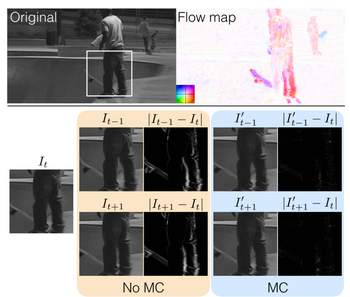


Spatial Transformer motion compensation

The motion compensation module learns to warp one frame onto another. The warping flow map is estimated in a coarse (c) and fine (f) stages.

$$I'_{t+1} = \mathcal{I}\{I_{t+1}(\Delta_{t+1})\} \qquad \theta^*_{\Delta,t+1} = \arg\min_{\theta_{\Delta,t+1}} \|I_t - I'_{t+1}\|_2^2 + \lambda \mathcal{H}(\partial_{x,y} \Delta_{t+1})$$





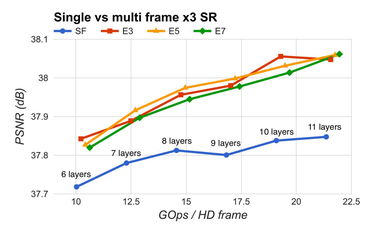
#ExperimentsAndResults

We use the CDVL dataset [4] containing 115 videos (1080p, 30fps) and train on sub-images of size 33x33 with Adam. Kernel size is 3. The number of features per layer is 24 in all cases. Computational efficiency is reported in number of floating point operations (GOps).

Spatio-temporal Networks (w/o motion compensation)

We find no gain in using more than 3 consecutive frames.

We also found early fusion (E) to be the best use of resources relative to slow-fusion (SF) designs.



Motion compensated video SR

Motion compensation corrects detailed structures compared to non-compensated networks.

State-of-the-art comparison

Variations of the proposed approach can improve accuracy (PSNR, SSIM), temporal consistency (MOVIE), and complexity (GOps).

	Image and video SR				Proposed VESPCN	
	Bicubic	SRCNN	ESPCN	VSRnet	5L-E3	9L-E3-MC
PSNR	25.38	26.56	26.97	26.64	27.05	**27.25**
SSIM	0.7613	0.8187	0.8364	0.8238	0.8388	**0.8447**
MOVIE ($\times 10^{-3}$)	5.36	3.58	3.22	3.50	3.12	**2.86**
GOps / 1080p frame	-	233.11	9.92	1108.73*	**7.96**	24.23

#Conclusion

We propose a network for motion compensation and video SR trainable end-to-end. This results in state-of-the-art accuracy and complexity, and temporally consistent reconstructions.

#References

[1] W. Shi et al., "Real-Time Single Image and Video Super-Resolution Using an Efficient", CVPR 2016.
[2] M. Jaderberg et al., "Spatial Transformer Networks", NIPS 2015.
[3] A. Kappeler et al., "Video super-resolution with convolutional neural networks", IEEE TCI 2016.