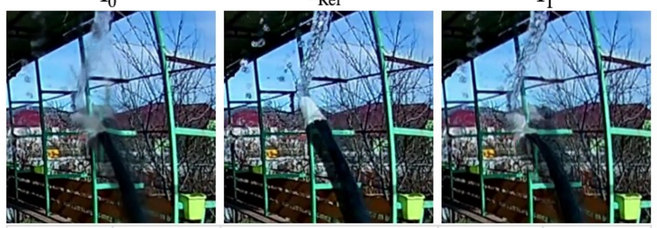
Shift-tolerant Perceptual Similarity Metric

Abhijay Ghildyal and Feng Liu

Portland State UNIVERSITY

Overview

Existing perceptual similarity metrics assume an image and its reference are well aligned. As a result, these metrics are often sensitive to a small alignment error that is imperceptible to the human eyes.



No-shift	1-pix-shift	Metric	No-shift	1-pix-shift
		Humans	✓	✓
✓		MS-SSIM		✓
✓		L2		✓
✓		LPIPS		✓

Motivation

As commonly expected, shifting one image by a few pixels will not alter human similarity judgment on a pair of images. We conducted a user study to verify this common belief.

Pixel shift	Number of user responses: Said Yes (Same)	Said No (Shifted)	Yes%	Avg. of std. in user responses per sample
0	140	10	93.3%	0.09 ± 0.17
1	121	29	80.7%	0.19 ± 0.23
2	84	66	56.0%	0.34 ± 0.21
3	52	98	34.7%	0.24 ± 0.23
4	52	98	34.7%	0.30 ± 0.24
5	40	110	26.7%	0.23 ± 0.24
6	35	115	23.3%	0.21 ± 0.24
7	31	119	20.7%	0.12 ± 0.20
8	27	123	18.0%	0.18 ± 0.23
9	15	135	10.0%	0.13 ± 0.21

Method

We build upon the Learned Perceptual Image Patch Similarity metric (LPIPS), a widely used learned metric, and explore architectural design considerations to make it robust against the imperceptible misalignment. Specifically, we study a wide spectrum of neural network elements, such as anti-aliasing filtering, pooling, striding, padding, and skip connection, and discuss their roles in making a robust metric.

Ablation Studies

We trained all our metrics using the original BAPPS training set on their original size of 256×256. We purposely did not train on the shifted version of the dataset to focus on neural network element designs.

- Combining BlurPool with reducing stride size makes the network more robust against imperceptible shifts (r_{rf}) and more consistent with human judgment (2AFC).
- Blur Before Activation works better when the stride size 1.
- BlurPool significantly improves the robustness of other backbones as well. A larger padding size improves 2AFC.

AA (BlurPool) Reflection-Pad 1	AA (BlurPool) Reflection-Pad 2	F-Conv	Stride in *conv-1*	2AFC	r_{rf} 1pixel	r_{rf} 2pixel	r_{rf} 3pixel
			4	70.65	2.87	3.92	3.74
✓			2	70.53	1.85	2.22	2.58
	✓		2	**70.67**	1.46	1.82	2.25
		✓	4	70.57	2.78	3.92	3.91
	✓	✓	2	70.52	1.77	2.15	2.48
			2	70.54	1.84	2.28	2.34
	✓		1	70.42	0.66	**1.13**	1.83
	✓	✓	1	70.44	**0.63**	1.14	**1.68**
✓§			2	70.57	2.63	3.36	3.16
	✓§		2	70.63	2.80	3.57	3.39
	✓§	✓	2	70.52	2.95	4.13	3.93

Anti-Alias (BlurPool)	Stride in *Conv-1*	BlurPool Location	2AFC	r_{rf} 1pixel	r_{rf} 2pixel	r_{rf} 3pixel
✓	2	Original	**70.67**	**1.46**	**1.82**	**2.25**
✓	2	FeatAfterBlur	70.55	1.73	1.84	2.49
✓	2	BlurBeforeAct	70.50	2.06	2.02	2.74
✓	1	Original	70.42	0.66	1.13	1.83
✓	1	FeatAfterBlur	**70.52**	0.69	1.11	1.60
✓	1	BlurBeforeAct	70.48	**0.57**	**1.06**	**1.50**

Network	AA (BlurPool) Reflection-Pad 1	AA (BlurPool) Reflection-Pad 2	2AFC	r_{rf} 1pixel	r_{rf} 2pixel	r_{rf} 3pixel
VGG-16			70.03	3.01	3.76	3.44
	✓		70.05	**0.66**	**1.08**	**1.44**
		✓	**70.07**	**0.66**	1.12	1.82
ResNet-18			69.86	2.67	3.35	3.77
	✓		69.95	**0.82**	**1.51**	**2.19**
		✓	**70.14**	1.07	1.81	2.38
Squeeze			69.61	7.41	7.58	10.35
	✓		69.24	**2.03**	3.06	3.93
		✓	**69.44**	2.10	**2.48**	**3.42**

Results

Our metrics are both more robust against imperceptible shifts and consistent with human visual similarity judgment than most of the similarity metrics on BAPPS dataset.

Network	2AFC	r_{rf} 1pixel	r_{rf} 2pixel	r_{rf} 3pixel
L2	62.92	3.59	7.55	10.82
SSIM [30]	61.41	3.16	7.20	13.73
CW-SSIM [31]	61.48	3.91	6.88	9.47
MS-SSIM [32]	62.54	2.22	5.83	10.66
PIEAPP Sparse [25]	64.20	2.83	3.19	3.81
PIEAPP Dense [25]	64.15	2.97	1.37	3.33
PIM-1 [3]	67.45	0.79	1.70	2.52
PIM-5 [3]	67.38	1.01	1.88	2.96
GTI-CNN [21]	63.87	3.95	4.91	7.88
DISTS [6]	68.83	2.85	2.89	4.03
E-LPIPS [16]	68.22	5.84	5.86	5.77
LPIPS (Alex) [37]	68.59	2.81	3.41	3.84
LPIPS (Alex) §∗†	70.54	2.58	3.59	3.53
LPIPS (Alex) ours∗†	70.39	0.66	1.24	1.79
LPIPS (Alex) §∗‡	**70.65**	2.87	3.92	3.74
LPIPS (Alex) ours∗‡	70.48	**0.57**	**1.06**	**1.50**

(§) Retrained from scratch. (∗) Trained on patches of size 256 using author's (†) / our (‡) setup.

Our method outperforms all other methods on the CLIC dataset while being shift-robust.

Network	Accuracy(%)	No. of rank flips 1pixel	2pixel	3pixel
L2	58.16	833	2102	2214
SSIM [30]	60.00	349	931	1109
PIEAPP [25]	75.44	91	134	158
E-LPIPS [16]	74.44	212	251	317
DISTS [6]	75.63	28	36	50
PIM-1 [3]	73.79	**13**	22	33
LPIPS(Alex) [37]	73.68	90	108	121
LPIPS(Alex)§∗†	76.53	59	51	62
LPIPS(Alex) ours∗†	**76.97**	17	**14**	**21**

(§) Retrained from scratch. (∗) Trained on image patches of size 64 using author's (†) setup.

Based on the responses from the user study our metric follows the sensitivity of human perception to pixel shifts more accurately.

Metric	JND mAP%
SSIM [30]	0.722
LPIPS (Alex) [37]	0.757
LPIPS (Alex) §∗†	0.740
LPIPS (Alex) **ours**∗†	0.771
LPIPS (VGG) [37]	0.770
LPIPS (VGG) §∗†	0.769
LPIPS (VGG) **ours**∗†	**0.775**
DISTS [6]	0.766
PIM-1 [3]	0.773

(§) Retrained from scratch. (∗) Trained on image patches of size 64 using author's (†) setup.

LPIPS (Alex)	0.072	0.316
LPIPS (Alex) ours	0.149	0.149
LPIPS (Alex)	0.001	0.353
LPIPS (Alex) ours	0.002	0.009