The Small Faces That Vex AI Surveillance Systems

Originally published at https://blog.metaphysic.ai/the-small-faces-that-vex-ai-surveillance-systems/ · March 16, 2023

The inability of surveillance systems to genuinely and accurately enhance tiny faces so that they are recognizable has been an internet meme for a long time, besides occupying a strong strand of research in the computer vision sector.

The recent revolution in GPU-driven image processing and synthetic data in the AI sector has given rise to hopes that simulated data could help train surveillance systems to recognize faces when they only occupy a small part of the footage; but at the same time, the assumption that such systems are 'reliable' stands on shaky legal ground, with some headline-grabbing failures having cropped up in recent years.

Available 'small in picture' resolutions for video surveillance systems, at varying image qualities. Source: https://arxiv.org/pdf/1805.11519.pdf

It's clear, therefore, that facial recognition systems that may be called into evidence in legal cases need sound and explainable methodologies before they're used as proof that a particular person is featured in a video.

Likewise, there is also a growing need in the AI VFX scene for systems that can automatically recognize faces in footage, even at acute angles, or if they are very small in the picture – so that actor-specific algorithms can be applied to the footage, or for purposes of automatic dataset generation for deepfake training sets – currently a relatively laborious and only semi-automated pursuit.

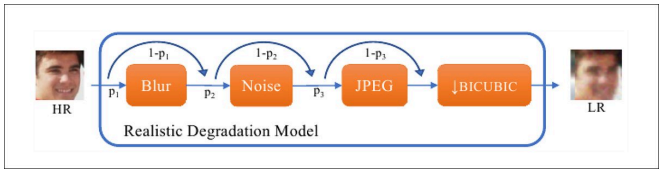
Thinking Small

A new paper from the Multimedia Signal Processing Group at Swiss Federal Institute of Technology (ETH Zurich) and École Polytechnique Fédérale de Lausanne (EPFL) proposes a solution: titled *Identity-Preserving Knowledge Distillation for Low-resolution Face Recognition*, the system developed for the project concentrates on the discriminative information *contained in low-resolution (LR) images*, rather than using the obtained features from such images and comparing them to the features obtained from high resolution (HR) images.

The latter technique could be called 'representative', and generates *LR>HR* mappings, or 'assumed' derived relationships between the LR features (which are quite indistinct and unreliable) and the HR features (which are very reliable but not necessarily an accurate analogue of their low-res counterparts, which means that people with similar feature profiles could mistakenly be identified as another person).EYH

Typical thinking in LR>HQ methodologies – that mapping a relationship between HR and LR features will create a reliable correlation – but it isn't necessarily so.

Among other innovations, the system relies on a 'data degradation model' capable of producing 'real world'-style low-res facial images – ones that are more authentic than prior methods, and more likely, once trained, to produce a recognition algorithm that will make a facial identification based on the low-res domain, rather than on unreliable relationships between the LR and HQ domains.

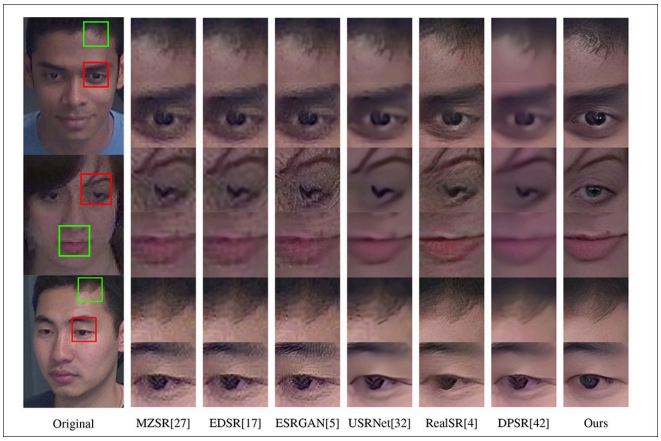


The pipeline for generating more authentic 'degraded' low-res images for training. The final stage, bicubic resizing, resizes the image using the least interpretative and aesthetically pleasing method available, ensuring that the image is not 'enhanced', which is undesirable when trying to ascertain accurate identifications. Source: https://arxiv.org/pdf/2303.08665.pdf

The data degradation pipeline that keeps the whole system focused on low-quality imagery, and which provides training data, is initially randomly corrupted by blurring operations, artificial noise, and then JPEG compression artefacts, before being downsampled by a crude and inelegant bicubic operation, which performs none of the smoothing or re-synthesis commonly found in standard server/platform-based frameworks such as ImageMagick, or in the increasingly sophisticated resampling algorithms found in professional tools such as Photoshop.

The authors' tests assert that the new method outperforms baseline models for the same task, though the authors concede that the system works best operating entirely within the LR domain.

This means (and not just for this project), that effective recognition systems of the future may need to adopt dual algorithms, with one of them specifically dedicated to *LR>LR* recognition – and this is a natural corollary of criticizing the assumed equivalency between *LR>HR* domains, which is barely suitable for AI-based image enlargement, and – it could be argued – entirely unsuited for recognition tasks.



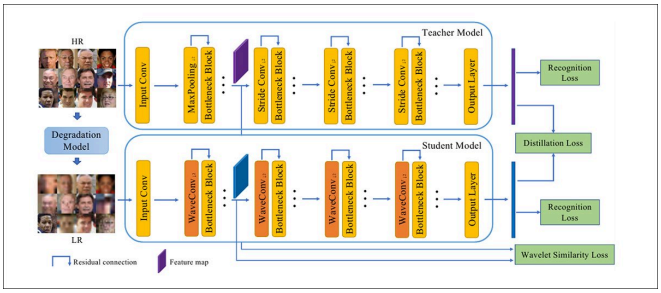
Varying results from diverse superresolution techniques, with varying losses of crucial identity information. Source: https://ietresearch.onlinelibrary.wiley.com/doi/pdfdirect/10.1049/ipr2.12359

Approach

The central problem that the new system must solve is that all 'gallery' or reference images which a recognition system will be using to attempt video surveillance identification are in themselves HQ images, whereas the available footage is likely to present a 'postage-stamp'-scale image of any face that might be a match.

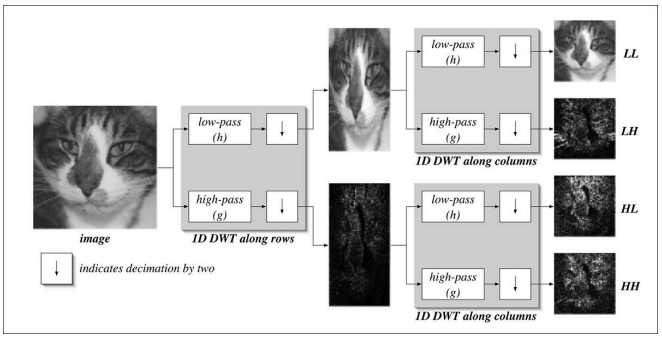
Besides this notable discrepancy in optical and resolution-based quality, issues such as obfuscation (the subject may be side-on to the camera, or have their face obscured) and video compression (video is typically delivered and stored compressed with various codecs, to lower stream latency and storage costs) can also affect the possible quality of surveillance-obtained faces.

The system proposed by the researchers uses a network that has been trained primarily on low-quality images, and where high-quality images have been deliberately degraded. Though the system is capable of comparing various resolutions of face images, it has been designed to concentrate on *LR><LR* matching, and, as mentioned, this is where it performs best.



The architecture for the new identity-preserving system.

The core of the architecture is an identity-preserving system called *WaveResNet*, which adapts ResNet by substituting the latter's pooling and stride-convolution layers with a low-pass filter built around Discrete Wavelet Transform (DWT).



The high-pass and low-pass layer workflow in DWT, which has been shoe-horned into ResNet for the purposes of the project. Source: https://mil.ufl.edu/nechyba/www/eel6562/course_materials/t5.wavelets/intro_dwt.pdf

The WaveResNet module effectively removes 'ambiguous' high-frequency information, and forces the registration of low-frequency features, constraining the system towards a 'minimal' LR gamut of features.

A Brilliant Student

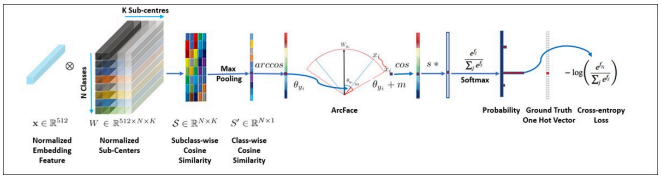
The architecture features a relatively typical teacher-student model knowledge distillation network, a predecessor of the combative generator/discriminator system in Generative Adversarial Networks (GANs), which successively improves the obtained losses between the supplied examples during training.

In a typical teacher-student model, the teacher model is relatively complex, whereas the student model is relatively simple. This effectively passes some of the burden of distillation (typically the work of a neural network that has no such conceptual limitations) into the design of the architecture itself.

This means that as such a network trains, it iterates through the supplied data, which may include real and synthetic (or altered/degraded) data, and slowly learns the relationships between the features obtained from the data samples; and that the more limited student network will *automatically* constrain itself to the most essential feature relationships.

However, for the new system, this is *not* the case: here, the student model contains the same number of parameters as the teacher model, and is, if anything, more capable; it is certainly more burdened: under this regime, the teacher model occupies itself exclusively with high resolution images, learning to extract rich features from them, before a process of *cross-resolution distillation* passes *multiple sizes* of those images on to the student network.

The central loss function for the new system is the 2020 ArcFace algorithm, which powers both the teacher and the student network. ArcFace features many of the facets (such as max pooling) substituted by the researchers when creating WaveResNet's variation on ResNet; but ArcFace offers these features under more proscribed conditions than in a base ResNet application.



Conceptual architecture for ArcFace. Source: https://arxiv.org/pdf/1801.07698.pdf

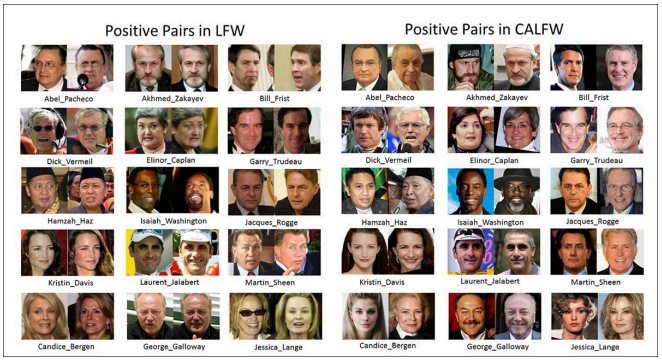
An additional loss function is used on the feature maps that are produced during this process, titled *Wavelet Similarity Loss*. This forces the student network to favor discriminative feature knowledge from the low-resolution space, helping to orient the entire system towards accurate recognition of low-quality images.

Since the new system 'downgrades' HR data, the final system operates most effectively on just the kind of tiny and poor-quality image that can be an obstacle in *LR>HR*-based systems.

Training and Tests

The authors tested their system using ArcFace's MSIM dataset, which comprises 3.28 million face images across 72,778 identities. All images were cropped to 112x122px resolution, and assigned five facial landmarks each (68 landmarks being the current standard for facial recognition). All images were downsampled to create HR/LR training pairs.

The system was evaluated on four datasets: LFW; AgeDB-30; CPLFW; and CALFW.



Example age pairs for single identities in CALFW. Source: https://arxiv.org/pdf/1708.08197.pdf

The teacher network was exclusively trained on high resolution images, and the student network on multiple obtained resolutions, all passed through the degradation model. Each network was trained for 18 epochs under the Standard Gradient Descent (SGD) optimizer at a batch size of 128, and at an initial learning rate of 0.1.

Methods					Avg on ∪{LFW,AgeDB,CPLFW,CALFW}				Overall Average
Backbone	Multi-scale	Degradation	KD	WaveSim	7x7	14x14	28x28	112x112(HR)	
ResNet					57.04	73.96	92.42	95.38	79.70
ResNet	✓				76.37	84.92	93.23	94.13	87.16
WaveResNet	✓				76.49	86.73	92.86	94.17	87.56
WaveResNet	✓	✓			76.95	87.77	92.55	93.30	87.64
WaveResNet	✓	✓	✓		75.41	88.31	93.18	94.31	87.80
WaveResNet	✓	✓	✓	✓	76.41	88.30	93.25	94.47	88.11

Results from the tests for the new system.

Of these results, the authors observe (as we mentioned earlier), that performance on HR images for the baseline models (listed in the results above) seems to deteriorate under this regime, but that the results on LR images seems to improve.

By inference, this suggests that if a single model cannot effectively straddle the HR and LR domains with equal accuracy, it may be necessary, pending further breakthroughs, to consider multiple identification systems running simultaneously and in concert, possible as two sub-networks of an overriding, explicit network, and silent 'handing off' tasks to each other as the input resolutions become larger or smaller, and therefore more apposite for one or the other of the networks.

The authors further comment:

'The rest of the results demonstrate the effectiveness of each proposed module. The proposed identity-preserving WaveResNet significantly improves the performance in LR testing data. Training with realistic synthetic data further improves the performance in LR scenarios but impairs recognition accuracy on HR images.

'The cross-resolution distillation framework remedies the performance sacrifice in HR images and improves the overall scores.

'Finally, after employing all proposed techniques, the model demonstrates promising results on both low and high-resolution images and outperforms the baseline models.'

Conclusion

Besides public video surveillance systems, the visual effects industry is going to increasingly need to distinguish multiple identities in footage, for a variety of reasons, and the new system proposed may be a step forward in achieving that aim more effectively – even if, at least for the time being, a difficult and stubborn problem may have to be addressed by dividing the workload across diverse networks, each capable of operating best within an LR or HR domain.