Vision Problems under Adverse Imaging Conditions

Zhu Li
Director, UMKC NSF Center for Big Learning
Dept of Computer Science & Electrical Engineering
University of Missouri, Kansas City
Email: zhu.li@ieee.org, lizhu@umkc.edu
Web: http://l.web.umkc.edu/lizhu
Short Bio:

Research Interests:

- Immersive visual communication: light field, point cloud and 360 video coding and low latency streaming
- Low Light, Res and Quality Image Understanding
- What DL can do for compression (intra, ibc, sr, inter, end2end)
- What compression can do for DL (compression, acceleration, distributed training)
Outline

- Overview of NSF CBL
- Research Overview
- Vision Problems under Adverse Imaging Conditions
  - Dark image enhancement from sensor field
  - Gradient image super resolution for key point repeatability
  - Human action recognition from RF signal domain
- Summary
NSF Industry-University Collaborative Research Center

• Who we are:
  • New *NSF Industry University Cooperative Research Center* established in 2018
    • $3M (NSF) + $3M (industry) investment
  • 30+ Faculties from 3 universities
  • 25 industry partners

• What we do:
  • First NSF national center on *big data/big learning*
  • Accelerate big data and AI's impact in the industry and society
  • UMKC: embedded deep learning for imaging, communication, IoT, and *medical applications*
Supporting Industry Members

- Remote Sensing and Hyperspectral Imaging
- 3D sensing, high precision 3D map for smart city and autonomous driving
- Robotics intelligence in cloud
- Image Processing and Understanding
- Data Mining and Database
- Financial Data Mining and AI
- Medical Imaging
- Speech recognition and synthesis

* Partial Collections
UMKC Faculties:

- Zhu Li (SCE): embedded deep learning for imaging, 3D sensing and communication
- Yugi Lee (SCE): data mining and deep learning/medical applications
- Praveen Rao (SCE): AI and database/big data applications
- Sejun Song (SCE): AI in networking, mobile, data centers, and IoT
- Hank Lee (MST): Medical Imaging
- Baek-Young Choi (SCE): AI in networking and IoT
- Reza Derakhshani (SCE): machine learning in biometrics/biomedical and imaging
- Chi Lee (Pharm): precision drug delivery
- Peter Koulen (Vision Research Center): human vision system

UMKC CBL Collaborators:

- George York (USAF Academy): UAV control and vision, imaging.
- Aggelos Katsaggelos (Northwestern Univ): leads Northwestern GAIM: Group for AI in Medicine, long time collaborator, will join CBL as a new member.
- Lingjia Liu (Virginia Tech): 5G wireless tech, mobile edge 3D sensing and point cloud services
- Ting Xie (Stowers Inst): Stem cell research, sub-micron accuracy retina imaging.
CBL 2019 Projects – Big Data Imaging, Auto Driving

Mobile Edge 3D Sensing and Point Cloud Services

Low Light Image Enhancement

QIK: Query Image via Knowledge Graph

sponsors:

1st year funding: $150K (NSF)+$200K (Industry)
Chroma Prediction in Future Video Coding

CBL 2019 Projects - Deep Learning in Coding

Immersive Media Coding
5 Year CBL Roadmap

**Business**
- Scene, Text, Action Vision Understanding
- NLP Understanding and Generation
- EHR, ICU, MRI, Genomics
- Multimodal Sensor Fusion

**Health**
- Virtual Assistant
- Cancer: breast, brain, lung; Sepsis; Behavior
- Wearables, Games, Living, Insurance
- Defense, surveillance, security, privacy

**IoT**
- Logistics, Marketing, Finance, Insurance
- Virtual Caregiver
- Real-time Diagnosis and Decision
- Intelligent City, Power, Transportation

**Applications**
- Multi-modal, cross-modal, hybrid: Vision, Text, Genomics, Business, IoT, Health
- CNN, RNN Variants
- Attention Memory
- Deep Bayesian
- Dynamic MetaDL
- Causality, Resilience
- GAN, Variational Multimodal, Mixture
- Spatial Temporal
- Knowledge Graph
- Never-ending Metacognition
- Hybrid Man-Machine Intelligence
- Fast, Hybrid DRL
- Supervised, Generative, Reinforcement
- Unsupervised, Never-ending, Hybrid

**Algorithms**
- DeepLite
- DeepCloud
- Intelligent Platform Design, Prototype, Deployment, and DevOps
- Intelligent Platform Design, Prototype, Deployment, and DevOps

**Systems**
- Deploy, Production, Massive Services
- Business
- Health
- IoT
- Security
- Refactors, Vertical Sectors, Transfer to Industry Partners, Other Campuses, and Community

ICT Strategic Sectors

ICT

Small Business or Traditional Companies

UMKC
Outline

- Overview of NSF CBL
- Research Overview
  - Vision Problems under Adverse Imaging Conditions
    - Dark image enhancement from sensor field
    - Gradient image super resolution for key point repeatability
    - Human action recognition from RF signal domain
  - Summary
Motivation

- Low-light photography
  - Almost all smartphone cameras have dedicated sections for low-light imaging
Motivation

- Low-light vision task
  - Object detection
  - Face recognition
  - Surveillance

Figure 2. Low light pedestrian detection (Ref: Multispectral Deep Neural Networks for Pedestrian Detection)

Figure 3. Low light pedestrian detection (Ref: Multispectral Deep Neural Networks for Pedestrian Detection)
Objective

- To design network to denoise the low-light image in Bayer domain
- To use wavelet decomposition to divide and conquer the problem by learning sensor field sub images using separate networks

Figure 4: [a] Extreme low-light image from Sony a7S II exposed for 1/25 second. [b] 250x intensity scaling of image in [a]. [c] Ground truth image captured with 10 second exposure time. [d] Output from SID[]. SID introduced some artifacts around the edge of the chair as shown by green arrow. [e] Output from ResLearning[]. The white region as indicated by arrow in image is not properly reconstructed as white compared to that in ground truth image. [f] Our result.
Under low-light condition image sensor suffers from low signal-to-noise ratio

Generates noisy image, as not enough photon reaches the camera sensors

Enlarging aperture will reduce the depth of field –blurry image

Extending the exposure time cause motion blur

Increasing the ISO will also amplify the noise signals

Figure 5. Effect of aperture, shutter speed and ISO in camera for low light imaging
Main Contribution

- Proposed a novel method of denoising before ISP (can be more useful for machine vision instead of human consumption)
- Decomposed the input raw image into low and high frequency subimages using wavelet transform
- A new loss function for learning high frequency components of our proposed wavelet decomposition network
Dataset

- See-In-Dark Dataset: Real world extreme low-light images with corresponding noise-free ground truth
- Illumination less than 0.5 lux
- Three different exposure of $1/10^{th}$, $1/25^{th}$ and $1/30^{th}$ seconds and corresponding ground truth of 10 seconds
- The time difference between the shutter speed is taken as the amplification ratio

Figure 6. Sample of low-light image and its corresponding ground truth image
Wavelet Decomposition

- Used Haar wavelet as decomposition filter
- $g(n)$ is low pass filter, $h(n)$ is high pass filter
- The resulting output is downsampled by half in rows and columns
- LL is equivalent to low freq while LH, HL and HH equivalent to horizontal, vertical and diagonal component respectively

Figure 7. One layer decomposition using wavelet transform.

Figure 8. Decomposition of image using wavelet transform. [a] Noisy low-light image converted to sRGB by using Rawpy library [b] Wavelet decomposition of small patch of [a]. [c] Wavelet decomposition of corresponding ground truth image [d] Prediction from our network for LL, LH, HL and HH component with combination of L1 and SSIM for high frequency component.
Methodology

- Two stages: first stage is the denoising network while the second stage is the off-the-shelf camera ISP
- Trained four different network for LL, LH, HL and HH component of wavelet
- Combined the information of LL to LH, HL and HH for better prediction of high frequency information

Figure 9: Overview of our wavelet decomposition based network. The first stage learns the decomposed image and used the inverse wavelet to reconstruct the denoised 4 channel image. The second stage uses the off-the-shelf ISP to enhances the image and converts into 3 channel sRGB image.
Network Architecture

- Network based on residual learning
- Consists of 32 residual blocks for LL while only 8 residual blocks were used for LH, HL and HH network
- LeakyReLU as activation function
- Residual block followed by Squeeze-and-Excitation block—converges the network faster and increases the performance
- While training, patch size is 256 x 256, learning rate of 0.0001, and 64 filters at each conv layer
- L1 as loss function, Adam as optimizer, and each network trained for 4000 epochs

Figure 10. (a) Residual network (b) Residual block with LeakyReLU as activation function and squeeze-and-excitation block
We use L1 loss for learning low frequency component (LL),

$$L_1 = ||\hat{x} - x||$$  \hspace{1cm} (1)

For high frequency component LH, HL and HH we used adaptive loss of L1 and SSIM loss

$$L_{structural} = 1 - SSIM(\hat{x}, x)$$  \hspace{1cm} (2)

$$L_{HF} = \alpha \cdot L_1 + L_{structural}$$  \hspace{1cm} (3)
Quality Metrics

- Evaluation against the current SOTA

Table 1. Comparison of our proposed method of denoising before ISP with the existing method of joint denoising and demosaicing. (Higher value of PSNR is better. Lower value of RMSE and NIQE is better.)

<table>
<thead>
<tr>
<th>Experiments</th>
<th>PSNR</th>
<th>RMSE</th>
<th>NIQE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SID[2]</td>
<td>28.97</td>
<td>0.03956</td>
<td>5.1904</td>
</tr>
<tr>
<td>ResLearning[1]</td>
<td>29.16</td>
<td>0.03926</td>
<td>5.8507</td>
</tr>
<tr>
<td>Ours</td>
<td>30.02</td>
<td>0.03568</td>
<td>4.6166</td>
</tr>
</tbody>
</table>

Table 2. Ablation study of our proposed method in terms of PSNR.

<table>
<thead>
<tr>
<th>Experiments</th>
<th>100x</th>
<th>250x</th>
<th>300x</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNet + ISP</td>
<td>31.95</td>
<td>29.49</td>
<td>27.64</td>
<td>29.52</td>
</tr>
<tr>
<td>ResLearning + ISP</td>
<td>32.25</td>
<td>29.70</td>
<td>27.70</td>
<td>29.70</td>
</tr>
<tr>
<td>Wavelet+L1+ISP</td>
<td>32.14</td>
<td>29.85</td>
<td>27.97</td>
<td>29.82</td>
</tr>
<tr>
<td>Wavelet+L1 &amp; SSIM + w/o data separation + ISP</td>
<td>32.10</td>
<td>29.61</td>
<td>28.22</td>
<td>29.84</td>
</tr>
<tr>
<td>Ours</td>
<td>32.34</td>
<td>29.97</td>
<td>28.22</td>
<td>30.02</td>
</tr>
</tbody>
</table>
Experimental Results

Figure 11. Results showing image details using our method in comparison to SID[] and ResLearning[]. [a, b] Dark input images [c, e] Outputs from SID[]. The text are blurred and color is different from ground truth. [d, f] Output from ResLearning. Though the image has lots of details than [c,e], the text is still blurred. [g, i] Outputs from our network. The text are much cleaner and color is much closer to the ground truth. [h, j] Zoomed version of corresponding ground truth images.
Experimental Results

Figure 12. [a] Subjective results from our method in comparison with BM3D[], SID[] and ResLearning[]. [a] Extreme low-light image captured by Sony a7S II. [b] Intensity scaled version of [a] converted to RGB by rawpy library [c] Denoised by BM3D and demosaic and enhanced by Rawpy library. We used different sigma values of 10, 20, 40, and 60 and selected the one with best PSNR. BM3D was not able to denoise properly as seen in the zoom image [d] Output from SID. We can see some artifacts indicated by arrow and bounding box [e] Output from ResLearning. The color reproduction is accurate. [f] Our result. Denoised in Bayer domain using wavelet decomposition and demosaic and enhanced by Rawpy library [g] Corresponding ground truth image.
More Results

Figure 13. Comparison of our method with BM3D[2], SID[1] and ResLearning[3] in terms of PSNR for the indoor image under extreme low-light condition. The color in the wall and the floor is well reproduced and closer to the ground truth image.
Figure 14. Comparison of our method with BM3D[2], SID[1] and ResLearning[3] in terms of PSNR for the outdoor image under extreme low-light condition. The detail in the image produced by our method is much closer to ground truth image.
Figure 15. Another example showing both color and details from our proposed method which is closer to the ground truth image. BM3D[2] uses the sigma value of 5. Though the texture is preserved, the color is different from output. SID[1] and ResLearning[3] have missing details and are blurred.
Conclusion and Future Work

- We propose a nobel method of direct sensor field denoising solution by exploiting the strong prior obtained from wavelet decomposition.
- We achieved significant gain in terms of PSNR via our decomposition network and loss function adaptation.
- The time complexity for our network is less than typical implementation, as we are processing approx two-third less information than sRGB image.
- Inference time is 21x faster (11 ms per 4K frame) than prior state of the art.
- In future, we will explore different wavelet functions, develop prefiltering and design adaptive loss function for even more performance gain.
Gradient Image and Multi-scale Representation

- **Gradient image** generally refers to a change in the direction of the intensity or color of an image. In a gradient image, in a certain direction, each pixel finds out the change in intensity of that same point in the original image.

- **Harris Detector** is used to find out the edges and extract corners of the image as well as discovering the infer features of the image.

- **Laplacian of Gaussian** is used for blob detection. It detects points that are continuously local maxima or minima with respect to both scale and space.

- In **SIFT**, difference of Gaussian (DoG) is used for feature detection. From DoG images, maxima and minima are computed to find key points in SIFT detection.
Proposed Method Formulation

- Let \( I(x, y) \) is the original image; \( G \) is the Gaussian Kernel,

\[
G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \tag{1}
\]

\[
L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{2}
\]

\( L \) is the function which denotes the scale space of the input image \( I \)

- Therefore Difference of Gaussian will be:

\[
D(x, y, \sigma_1, \sigma_2) = (G_1(x, y, \sigma_1) - G_2(x, y, \sigma_2)) * I(x, y) \tag{3}
\]

\[
D(x, y, \sigma_1, \sigma_2) = L_1(x, y, \sigma_1) - L_2(x, y, \sigma_2) \tag{4}
\]

- The standard deviation values, \( \sigma \) are 1.24, 1.54, 1.94, 2.45, 3.09 for formulating 4 different DoGs
Proposed Method Formulation

- The loss function $E$ is the MSE loss between the DoG of the super-resolved blurred generated image and the DoG from convolution with original image:

$$E(\hat{D}, D_{original}) = \sum_{i=1}^{n} \sum_{j=1}^{m} (\hat{D}^{ij} - D_{original}^{ij})^2 \quad (5)$$

Where $\hat{D}$ is the predicted DoG image which is upsampled and $D_{original}$ is the DoG image computed from the original one convolved with Gaussian filter.

- The gradient descent of the loss function will be:

$$\frac{\delta E}{\delta \hat{D}} = \frac{\delta(\sum_{i=1}^{n} \sum_{j=1}^{m} (\hat{D}^{ij} - D_{original}^{ij})^2)}{\delta \hat{D}} \quad (6)$$

\begin{align*}
\frac{\delta E}{\delta \hat{D}} &= 2 \sum_{i=1}^{n} \sum_{j=1}^{m} (\hat{D}^{ij} - (\frac{1}{2\pi\sigma^2} P - \frac{1}{2\pi\sigma^2} Q)) \quad (7) \\
&\quad (1 - (\frac{1}{2\pi\sigma^2} \frac{\delta P}{\delta \hat{D}} - \frac{1}{2\pi\sigma^2} \frac{\delta Q}{\delta \hat{D}}))
\end{align*}

$$P = e^{\frac{-x^2+y^2}{2\sigma^2}} \ast I(x_i, y_j), \quad Q = e^{\frac{-x^2+y^2}{2\sigma^2}} \ast I(x_i, y_j) \quad (8)$$

- The simplified loss function can be written as MSE between Gaussian blurred images and computing DoG images separately.

$$E(\hat{L}, L_{original}) = \sum_{i=1}^{n} \sum_{j=1}^{m} (\hat{L}^{ij} - L_{original}^{ij})^2 \quad (9)$$
Network Implementation

- Low Resolution input images will be passed through a deep learning based gradient image super resolution stage. There are five SR networks for the purpose.
- Each SR network produces a super-resolved Gaussian blurred image with different $\sigma$ values \([\sigma = \{1.24, 1.54, 1.94, 2.45, 3.09\}]\)
- Four Gradient images (DoG image) are computed from five Gaussian Blurred images.
- Four Gradient images are integrated to SIFT method for the computation of key matching points.

Figure: Proposed Network Architecture
Network Implementation

**Figure**: Deep learning Gradient Image Super resolving network to compute upscaled gradient image

- **Filter kernel size of 3X3 with 64 number of features**
- **Deconvolutional Layer is used to upscale.**

**Figure**: Residual Blocks
Alternative Network Implementation

- Filter kernel size of 3X3 with 64 number of features
- Deconvolutional Layer is used to upscale.

**Figure:** Deep learning Gradient Image Super resolving network to compute upscaled gradient image

**Figure:** Residual Blocks
Experimental Dataset

Training Dataset:
1. CVPR DIV_2k dataset with 800 images is used for training.
2. They are first downsampling by 2/4 times.
4. Total input data 300k.

Test Dataset:
1. MPEG CDVS Full dataset.
2. MPEG CDVS is a comprehensive collection of images of various objects which consists of 186k labeled images of CDs and book covers, paintings, video frames, buildings and common objects.
5. 200 matching pairs from each category were chosen.
6. They are first downsampling by 2/4 times.
## Results

**MPEG CDVS Full dataset results:**

**Table 1:** Average number of SIFT matching points for 200 matching image pairs from each category

<table>
<thead>
<tr>
<th>Category</th>
<th>Upscaling Factor</th>
<th>Avg. no. of matching SIFT points for the original image</th>
<th>Avg. no. of matching SIFT points using proposed method-1</th>
<th>Avg. no. of matching SIFT points using proposed method-2</th>
<th>Avg. no. of matching SIFT points using EDSR</th>
<th>Avg. no. of matching SIFT points using SRCNN</th>
<th>Avg. no. of matching SIFT points using SRGAN</th>
<th>Avg. no. of matching SIFT points using bicubic interpolation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building</td>
<td>2</td>
<td>125.8</td>
<td>124.5</td>
<td>130.4</td>
<td>116.3</td>
<td>114.5</td>
<td>115.8</td>
<td>112.4</td>
</tr>
<tr>
<td>Building</td>
<td>4</td>
<td>125.8</td>
<td>110.8</td>
<td>115.4</td>
<td>105.6</td>
<td>104.2</td>
<td>104.3</td>
<td>100.4</td>
</tr>
<tr>
<td>Graphics</td>
<td>2</td>
<td>101.6</td>
<td>99.8</td>
<td>102.8</td>
<td>94.5</td>
<td>93.8</td>
<td>94.2</td>
<td>92.8</td>
</tr>
<tr>
<td>Graphics</td>
<td>4</td>
<td>101.6</td>
<td>87.2</td>
<td>90.4</td>
<td>86.7</td>
<td>86.1</td>
<td>86.8</td>
<td>85.4</td>
</tr>
<tr>
<td>Objects</td>
<td>2</td>
<td>115.3</td>
<td>113.9</td>
<td>118.5</td>
<td>106.9</td>
<td>103.9</td>
<td>104.8</td>
<td>102.6</td>
</tr>
<tr>
<td>Objects</td>
<td>4</td>
<td>115.3</td>
<td>105.1</td>
<td>108.8</td>
<td>99.1</td>
<td>98.2</td>
<td>98.5</td>
<td>96.2</td>
</tr>
<tr>
<td>Painting</td>
<td>2</td>
<td>114.4</td>
<td>114.7</td>
<td>120.5</td>
<td>105.9</td>
<td>104.4</td>
<td>104.9</td>
<td>100.7</td>
</tr>
<tr>
<td>Painting</td>
<td>4</td>
<td>114.4</td>
<td>106.1</td>
<td>109.8</td>
<td>101.5</td>
<td>100.1</td>
<td>100.2</td>
<td>96.1</td>
</tr>
<tr>
<td>Video</td>
<td>2</td>
<td>94.3</td>
<td>90.3</td>
<td>94.4</td>
<td>87.2</td>
<td>86.2</td>
<td>85.8</td>
<td>85.2</td>
</tr>
<tr>
<td>Video</td>
<td>4</td>
<td>94.3</td>
<td>82.2</td>
<td>85.5</td>
<td>80.1</td>
<td>79.4</td>
<td>79.6</td>
<td>79.2</td>
</tr>
</tbody>
</table>
## Results

### Oxford dataset results: Average number of SIFT matching points for 200 matching image pairs

<table>
<thead>
<tr>
<th>Upscaling Factor</th>
<th>Avg. no. of matching SIFT points for the original image</th>
<th>Avg. no. of matching SIFT points using proposed method-1</th>
<th>Avg. no. of matching SIFT points using proposed method-2</th>
<th>Avg. no. of matching SIFT points using EDSR</th>
<th>Avg. no. of matching SIFT points using SRCNN</th>
<th>Avg. no. of matching SIFT points using SRGAN</th>
<th>Avg. no. of matching SIFT points using bi-cubic interpolation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>105.4</td>
<td>101.4</td>
<td>107.3</td>
<td>97.1</td>
<td>96.2</td>
<td>96.4</td>
<td>94.2</td>
</tr>
<tr>
<td>4</td>
<td>105.4</td>
<td>93.2</td>
<td>97.8</td>
<td>91.1</td>
<td>90.4</td>
<td>90.3</td>
<td>89.9</td>
</tr>
</tbody>
</table>

### Paris dataset results: Average number of SIFT matching points for 200 matching image pairs

<table>
<thead>
<tr>
<th>Upscaling Factor</th>
<th>Avg. no. of matching SIFT points for the original image</th>
<th>Avg. no. of matching SIFT points using proposed method-1</th>
<th>Avg. no. of matching SIFT points using proposed method-2</th>
<th>Avg. no. of matching SIFT points using EDSR</th>
<th>Avg. no. of matching SIFT points using SRCNN</th>
<th>Avg. no. of matching SIFT points using SRGAN</th>
<th>Avg. no. of matching SIFT points using bi-cubic interpolation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>110.5</td>
<td>107.9</td>
<td>113.2</td>
<td>101.4</td>
<td>99.2</td>
<td>99.8</td>
<td>99.1</td>
</tr>
<tr>
<td>4</td>
<td>110.5</td>
<td>99.8</td>
<td>102.4</td>
<td>97.1</td>
<td>95.4</td>
<td>95.9</td>
<td>95.3</td>
</tr>
</tbody>
</table>
Comparative Results for SIFT matching points

(a) SIFT matching points for original image (102 points)  
(b) SIFT matching points using proposed method (112 points)

(c) SIFT matching points using EDSR (100 points)  
(d) SIFT matching points using bicubic interpolation (96 points)

Figure: SIFT Matching Points Comparison for a sample matching image pair with 2x upscaling
Privacy-Preserving Fall Detection with Deep Learning on mmWave Radar Signal
Outline

- Introduction
- Framework
- Radar Signal Processing
- Experimental Devices
- Network
- Experimental Results
Introduction

- Fall injuries lead the accidental death and nearly $34 billion in direct medical costs annually for seniors.

- Conventional solutions:
  - Wearable portable alert devices, e.g. automatic bracelets.
    - Pros: Accuracy and low-latency
    - Cons: Skin discomfort and inconvenience
  - Nonwearable alert system, e.g. camera-based surveillance equipment.
    - Pros: Accuracy and low-latency
    - Cons: High power consumption, invasion of privacy and high sensivenes at extreme environment

- Related works:
  - Doppler based radar detection [1]
  - The changes of different WiFi channel solution [2]
  - 3D-CNN radar frequency detection [3]

---

Motivated by the 3D-CNN RF-based solution, we propose an LSTM-based fall detection method based on the mmWave radar signal.

- Characterize the radar reflections based on distance from the human body along with the vertical and horizontal angles of arrays.
- Capture locality and velocity components simultaneously.
- Radar signal low-dimension embedding algorithm (RLDE) with LSTM reduces the complex and save chip memory.

![Figure 1. mmWave Radar based Fall Detector](image-url)
Proposed radar signal-based fall detection

- Human activities are regarded as the changes in terms of range, angle, and speed, which can be caught by a pair of IWR1642 radar devices.

- The time interval and intensity of signal between the receiver (RX) and transmitter (TX) can be recorded and correlated to fundamental attributes by training.

- The proposed method comprises two subtasks:
  - Radar signal processing
  - Neural network processing

Figure 2. Framework of Proposed Detector
This procedure performs the frequency modulated continuous wave (FMCW) signal conversion to analyzable digital form in the spatial domain (reflection heatmaps).

- ADC (Analog-digital converter): modulate continuous form to discrete form.
- Range-FFT (Range domain Fast Fourier Transform): convert the signal from the time domain to the spatial (range) domain.
- Angle-FFT (Angle domain Fast Fourier Transform): catch phase difference between each RX antenna.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. Range</td>
<td>10 m</td>
<td>Wave Form</td>
<td>FMCW</td>
</tr>
<tr>
<td>Range Res.</td>
<td>4 cm</td>
<td>Frequency</td>
<td>77-81 GHz</td>
</tr>
<tr>
<td>Num of RX</td>
<td>8</td>
<td>Num of TX</td>
<td>4</td>
</tr>
<tr>
<td>Field of View</td>
<td>120°</td>
<td>Angular Res.</td>
<td>15°</td>
</tr>
<tr>
<td>Max. Velocity</td>
<td>6.5 m/s</td>
<td>ADC samples</td>
<td>256</td>
</tr>
<tr>
<td>Velocity Res.</td>
<td>0.2 m/s</td>
<td>Frame rate</td>
<td>25 f/s</td>
</tr>
<tr>
<td>Wavelength</td>
<td>3.9 mm</td>
<td>Max. Bandwidth</td>
<td>3,750 MHz</td>
</tr>
</tbody>
</table>
Human activities are continuous dynamic patterns that can be recognized in both spatial and temporal dependencies. We use successive radar reflection heatmaps as the representative of human activities.

- PCA is adopted as RLDE algorithm to project reflection heatmaps \( \{H_t, V_t\} \) to a low-dimension subspace \( P \) as the elimination of spatial redundancies,
- The proposed RNN with LSTM units utilizes the changes of motion at the temporal domain. The softmax layer operates as a classifier. The cross-entropy function is adopted as the objective function.

**Figure 3. Architecture of RNN with LSTM units**
4,126 samples (2.56s for each sample) consist of 128 frames of reflection heatmaps, divided into two classes: fall and non-fall.

Table 1: The comparison on accuracy and processing time between 3DCNN and LSTM with or w/o RLDE implementation

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Training time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o RLDE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3DCNN</td>
<td>95.3%</td>
<td>96.6%</td>
<td>96.0%</td>
<td>181.21</td>
</tr>
<tr>
<td>LSTM</td>
<td>100.0%</td>
<td>93.6%</td>
<td>96.7%</td>
<td>94.29</td>
</tr>
<tr>
<td>LSTM^64</td>
<td>100.0%</td>
<td>97.9%</td>
<td>98.9%</td>
<td>56.83</td>
</tr>
<tr>
<td>LSTM^32</td>
<td>100.0%</td>
<td>95.8%</td>
<td>97.8%</td>
<td>37.22</td>
</tr>
<tr>
<td>LSTM^16</td>
<td>100.0%</td>
<td>97.7%</td>
<td>98.9%</td>
<td>22.21</td>
</tr>
<tr>
<td>LSTM^8</td>
<td>97.9%</td>
<td>100.0%</td>
<td>98.9%</td>
<td>20.33</td>
</tr>
<tr>
<td>LSTM^4</td>
<td>100.0%</td>
<td>97.7%</td>
<td>98.9%</td>
<td>17.08</td>
</tr>
<tr>
<td>LSTM^2</td>
<td>97.5%</td>
<td>88.6%</td>
<td>92.9%</td>
<td>15.12</td>
</tr>
<tr>
<td>with RLDE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Facebook Connectivity Workshop, 2018
Multiple human activities detections: 7 categories of human activities are labeled: Boxing, Falling, Jogging, Jump, Pick up, Stand up & Walking.

**Figure 4.** Accuracy of Multiple Human Activities Detecting

**Confusion Matrix of Multiple Human Activities**

<table>
<thead>
<tr>
<th>True Class</th>
<th>Boxing</th>
<th>Falling</th>
<th>Jogging</th>
<th>Jump</th>
<th>Pickup</th>
<th>Standup</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boxing</td>
<td>97.7%</td>
<td>2.3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Falling</td>
<td>1.2%</td>
<td>69.4%</td>
<td>1.2%</td>
<td>1.2%</td>
<td>3.5%</td>
<td>15.3%</td>
<td>8.2%</td>
</tr>
<tr>
<td>Jogging</td>
<td></td>
<td>100.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jump</td>
<td>1.8%</td>
<td>96.4%</td>
<td></td>
<td>1.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pickup</td>
<td>5.9%</td>
<td>91.2%</td>
<td>2.9%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standup</td>
<td>32.1%</td>
<td>5.7%</td>
<td>49.1%</td>
<td>13.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Walking</td>
<td></td>
<td>0.7%</td>
<td>99.3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Average Inference Time Complexity:**
- RLDE + LSTM: 0.06042 sec
- 3DCNN: 7.336 sec
Conclusion & Future Work

☑ Summary

▪ Radar signal domain contains enough info for a variety of vision tasks, while have the feature of privacy preserving
▪ Introducing deep learning schemes with rich prior constraints of radar signal can potentially achieve better performances
▪ This is an initial work that shows promising results

☑ Future Work

▪ Larger data set with richer and fine granular labeling of human actions automatically and semiautomatically from cameras
▪ Potential compressive sensing + deep learning to by-pass the radar signal processing pipeline after ADC
Thank You