SIT789 – Robotics, Computer Vision and Speech Processing

High Distinction Task 5.2: Minor research project

Objectives

The objectives of this lab include:

- Reading and understanding a research paper in Computer Vision/ Speech Processing
- Performing literature review
- Programming upskilling
- Getting some experience on technical writing

Tasks

You are given a list of research papers with references for implementation (see the table below). This is individual task and thus each student will need to select a paper from the list. However, you are always welcome to recommend other papers/topics. In case that you have your own papers and would like to use them for this task, you should consult with the Unit Chair to make sure your recommended papers align with the Unit.

After choosing a paper, you need to read and understand the paper. Note that the topic of the paper may not be covered in the lectures. Therefore, to understand a paper, you may need to read other related papers and/or do further research. This is an opportunity for you to expand your knowledge and get updated with the literature of the chosen topic.

You also need to run the code implementing your selected paper. Note that all the papers in the list are accompanied with code. However, you are free to search and use other implementations. If you want to use your own paper, make sure that the paper's code is available, or you can re-implement it.

Your tasks include,

1) Report writing (**6 marks)**: Summarising your selected paper in a report. Your report should satisfy the following criteria.
    a. The report should reasonably summarise the paper/topic chosen
    - The report should be comprehensive and provide enough details (2 marks)
    - Technical information is presented accurately (2 marks)
    b. The report should have a good presentation
    - The report should have a reasonable layout; students are free to choose any layout and do not necessarily need to follow the layout of the paper chosen (1 mark)
    - The report should have a reasonable length, from 2,000 – 2,500 words depending on the paper/topic selected (0.5 marks)
    c. References (0.5 marks)
    - The paper should provide a list of references. Students can choose any style for the references (e.g., ACM, IEEE, etc.). However, all the references should be presented consistently in one style (e.g., the references should not mix-up both ACM and IEEE).
    - The references may not be exactly same as those of the paper chosen but should cover enough important work related to the paper/topic chosen.

2) Demonstration (**4 marks**).
    a. You need to demonstrate the chosen paper in both phases: training (if possible) and testing, including:
    - Collect and label a small dataset (1 mark)
    - Train and test the method presented in your chosen paper on the collected data (2 marks)

    *Note: It may not be always easy to collect and label a dataset that is large enough to train a model in several weeks. If you cannot make such a dataset, you can re-use some public datasets (even datasets used in your chosen paper). However, you should customise these datasets, e.g., changing the number of classes for image classification or detection problem, changing the number of samples used for training and testing. You are always welcome to contact your Unit chair to discuss about your chosen paper and expected outcomes.

    b. You need to report several qualitative results (i.e., visual effects, e.g., images). Note that you should NOT simply copy results shown in the paper. You should also describe the results and draw conclusions based on your own observations (1 mark)
    c. There is always room for improving/enhancing methods proposed in the paper (e.g., improving the accuracy, enhancing the speed). Therefore, it would be great if you could identify limitations of the work presented in the paper and suggest improvements. Note that, your proposed improvements should be specific and clear enough to be implemented though you are not required to implement them. Your proposed solutions should also be relevant to the Unit, e.g. improvements made by using more powerful hardware are not relevant to the Unit. Note that this part (c) is optional, but you may be granted some bonus score (1 mark).

To be considered for the HD level, you need to achieve at least 8 marks overall with no lower than 4 marks for the report and no lower than 3 marks for the demonstration. You are also required to attend an interview (either online or in-person depending whether you enrolled the Unit online or on-campus).

You can find reading and writing tips supplied in OnTrack. Please read through these slides as they could be useful for you to undertake this project.

Submission instructions

1. Submit your report and results to OnTrack.
2. Submit your code to OnTrack.

Papers	Topics	Source code	Notes
Computer Vision			
Vision transformer	Image recognition	Code	This work applies transformer, a recently popular technique.
YOLO-World	Object detection	Code	Since the first version, there have been many variants of YOLO developed. If you choose to work on this topic, you will need to perform a literature view of the YOLO's family.
YOLO-E	Object detection	Code	Since the first version, there have been many variants of YOLO developed. If you choose to work on this topic, you will need to perform a literature view of the YOLO's family.
SAM/SAM2	Object segmentation	Code (SAM, SAM2)	You can choose either SAM or SAM2. However, if you choose SAM2, you still need to understand SAM.
Realtime multi-person 2D pose estimation using part affinity fields	Human pose estimation	Code (Python, C++, Matlab)	
NVIDIA semantic segmentation	Semantic segmentation	Code	This work applies attention mechanism, a recently popular technique.
Deeplab	Semantic segmentation	Code (tensorflow), Pytorch re-implementation, Pytorch-vision	
U-Net	Semantic segmentation	Pytorch re-implementation	
YOLACT	Object/instance segmentation	Code	
Interactive segmentation	Object/instance segmentation	Code	
EAST	Text detection	Pytorch re-implementation, tensorflow re-implementation	
Ultrafast lane detection	Lane (road) detection	Code	
LaneATT	Lane (road) detection	Code	
RAFT	Optical flow estimation	Code	
HR-Depth	Depth estimation	Code	
Depth-VO-Feat	Depth estimation	Code	
DORN	Depth estimation	Code	
PatchNet	3D Object detection (3D Vision)	Code	Outdoor scenes
ImVoteNet	3D Object detection (3D Vision)	Code	Indoor scenes
Cylinder3D	Point cloud semantic segmentation (3D vision)	Code	
JSIS3D	Point cloud semantic and instance segmentation (3D Vision)	Code	You may experience with timeout issue if you run the code on Google Colab.
Pixel2Pixel	Image synthesis (image generation)	Code	This work requires some knowledge in generative adversarial networks (GAN).
CycleGAN	Image synthesis (image generation)	Code	This work requires some knowledge in generative adversarial networks (GAN).

StyleGAN	Image synthesis (image generation)	Code	This work requires some knowledge in generative adversarial networks (GAN)
DatasetGAN	Image synthesis (image generation)	Code	This work requires some knowledge in generative adversarial networks (GAN)
BachGAN	Image synthesis (image generation)	Code	This work requires some knowledge in generative adversarial networks (GAN)
NSD	Image synthesis (image generation)	Code	This work requires some knowledge in generative adversarial networks (GAN)
PyTorchVideo	Action recognition (video understanding)	Code	Not recommended for Windows
RFNet-4D	4D reconstruction	Code	
Speech Processing			
DeepSpeech: scaling up end-to-end speech recognition	Speech recognition	Re-implementation	
SpecAugment: a simple data augmentation method for automatic speech recognition	Speech recognition	Re-implementation	
Wav2vec: unsupervised pre-training for speech recognition	Speech recognition	Code	
Jasper: an end-to-end convolutional neural acoustic model	Speech recognition	Code	
Speaker recognition from raw waveform with sincnet	Speaker recognition/verification	Code	
ASV-CM reinforce	Speaker recognition/verification	Code	This work applies reinforcement learning
Kaldi-Pytorch speaker embedding	Speaker recognition/verification		You may experience with some difficulties in installing required software packages, e.g., Kaldi
3-D convolutional recurrent neural networks with attention model for speech emotion recognition	Speech emotion recognition	Code	This work applies attention mechanism, a recently popular technique
An interaction-aware attention network for speech emotion recognition in spoken dialogs	Speech emotion recognition	Code	This work applies attention mechanism, a recently popular technique
Joint cross-attentional A-V fusion	Visual-Audio emotion recognition	Code	This work combines both visual and speech signals for emotion recognition
NELE-GAN	Speech intelligibility enhancement	Code	
iMetricGAN	Speech intelligibility enhancement	Code	
Synthetic speech detection	Synthetic speech detection (Speech synthesis)	Code	
VQ-VAE	Speech reconstruction (Speech synthesis)	Code	Japanese/Chinese speech databases

Speech waveform modelling	Speech waveform modelling (Speech synthesis)	Code	
VQMIVC	Voice conversion (Speech synthesis)	Code	
Assem-vc	Voice conversion (Speech synthesis)	Code	
Multi-speaker text-to-speech	Text-to-speech (Speech synthesis)	Code	This work applies zero-shot adaptation
WaveRNN	Text-to-speech (Speech synthesis)	Re-implementation	The code repository includes the code/model for several papers related to the WaveRNN's topic
NISQA	Speech quality and naturalness assessment	Code	The code repository includes the code/model for several papers related to the NISQA's topic