Computer Vision EXAM

The exam can be taken in one of two ways:

1. by taking an **oral** exam on the topics covered during the course. This method follows the scheduled exam dates.
2. by completing a **project** according to the methods outlined below.

In both cases registering for an exam session is necessary to record the grade.

Project modalities

- The project deadline is independent of the scheduled dates for the oral exams. If you decide to take the exam with a project, please contact the professor to agree on the project to be carried out.
- The project must be submitted within a year from the date of assignment (after this period, please contact the professor to update the proposal).
- The project can be carried out by 1 or 2 people. In the second case, a greater analysis and in-depth investigation is expected (e.g. comparing different techniques).
- Those who intend to pursue this path are required to produce:
  - functioning code with adequate experimentation and analysis.
  - A report that frames the context, formally describes the methodology pursued, reports experimental results, and their analysis. The report can be written in either Italian or English of your choice.
  - Finally, the project must be presented (approximately one week after submitting the code and report) with a presentation of about 20 minutes (using slides if desired). Pay attention: during the presentation, questions may be asked to assess how much the concepts presented in class have been studied.
- Below are some project proposals. In case all these projects are assigned, the professor will propose other topics.

IF YOU ARE INTERESTED IN ONE OF THIS PROJECT CONTACT ME TO ARRANGE THE JOB

Project proposal

1) Video Fake detection

(This work is in conjunction with the department of Psychology, Università Cattolica. To coordinate the work on it, choose this if you are interested AND if you are planning to tackle it before summer break)

Study Description

This experimental study aims to compare videos of real people giving a speech with those generated by generative artificial intelligence (AI). Each video presents the same content, using the same text and the same duration, but it will not be revealed to the experimental subjects whether the video is authentic or produced by AI. The experiment could be repeated using only the audio tracks to verify if the absence of visual components influences the subjects' perceptions. Evaluation will including:

- Quantitative evaluation
- Perceptual evaluation
- Psychological evaluation

Final Objective of the Study

The objective of this study is to evaluate the extent to which AI-generated videos can effectively simulate human performance in terms of verbal and non-verbal communication. It aims to explore whether viewers are able to distinguish between authentic speeches and those artificially generated, and how various aspects of the video influence this perception.

Methodology:

- Produce real videos (according a give protocol)
- Produce fake videos using one of this solution (https://www.synthesia.io/pricing-options#anchor-yearly-table, https://app.heygen.com/avatars, https://app.dupdub.com/pricing)
- Produce quantitative comparison implementing a method such as the one described in “Bursic, S., D'Amelio, A., Granato, M., Grossi, G., & Lanzarotti, R. (2021, January). A quantitative evaluation framework of video de-identification methods. In 2020 25th international conference on pattern recognition (ICPR) (pp. 6089-6095). IEEE.”
- Produce a perceptive evaluation (Cartella, G., Cuculo, V., Cornia, M., & Cucchiara, R. (2024). Unveiling the Truth: Exploring Human Gaze Patterns in Fake Images. IEEE Signal Processing Letters.)
- Procedure of psychological evaluation(after viewing, participants will be asked to evaluate the videos on various aspects such as the interest aroused, level of engagement, clarity of presentation, and video and audio quality. Additionally, they will have to indicate whether they believe the video is real or fake.)

2) Speech emotion recognition - raw data (GIA' ASSEGNATO)

Study Description

This experimental study aims to explore the effectiveness of deep neural networks when confronted with limited data. Specifically, the project endeavors to analyze a selection of speeches encompassing eight distinct emotional states, sourced from the RAVDESS dataset. The analysis entails processing the raw data, namely the spectrogram or mel-spectrogram of the audio files, utilizing neural networks for feature extraction. Following the data-driven feature extraction phase, a final classification is executed employing a neural network architecture. It is imperative to note that the study prohibits the utilization of pre-trained models, as the primary objective is to assess the depth achievable with a restricted dataset.

Final Objective of the Study

The primary aim of this investigation is to assess the efficacy of deep neural networks in accurately recognizing emotions in speech, particularly under conditions of limited dataset availability. Additionally, a significant aspect of this study involves the exploration of an optimized architecture for the classifier.

Methodology:

- Extract the raw audio
- Extract significant features automatically with deep learning approaches
- Investigate on the most efficient architecture for speech emotion classification on the extracted features
- Compare the results obtained with state-of-the-art results (https://paperswithcode.com/sota/speech-emotion-recognition-on-ravdess)

- **To obtain the RAVDESS dataset from our server:**

1) Open a terminal and type: sftp student@turing.di.unimi.it (if you are asked about key/fingerprints, just type yes)

2) Insert the password: comelico39

3) move to the datasets/ directory: cd datasets/
4) download the dataset: get -r RAVDESS/ /path-to-your-local-directory/)

3) Speech emotion recognition - audio features (GIA' ASSEGNATO)

Study Description

This experimental study aims to explore the effectiveness of neural networks working on features theoretically tailored for speech recognition, when confronted with limited data. In fact, the lack of data is one of the main obstacles of the massive use of deep learning. Specifically, the project endeavors to analyze a selection of speeches encompassing eight distinct emotional states, sourced from the RAVDESS dataset. The analysis entails processing the raw data, namely the spectrogram or mel-spectrogram of the audio files, with standard audio procedures to extract relevant information. Following the feature extraction phase, a final classification is executed employing a neural network architecture. This network architecture should be optimized depending on the extracted features.

Final Objective of the Study

The primary aim of this investigation is to assess the efficacy of neural networks on hand-made features in accurately recognizing emotions in speech, particularly under conditions of limited dataset availability. Additionally, a significant aspect of this study involves the exploration of an optimized architecture for the classifier.

Methodology:

- Extract the raw audio
- Extract significant features with classic audio techniques (for instance, MFCC)
- Investigate on the most efficient architecture for speech emotion classification on the extracted features
- Compare the results obtained with state-of-the-art results (https://paperswithcode.com/sota/speech-emotion-recognition-on-ravdess)

4) 3D Face dataset preparation

Study Description

3D acquisition devices are used to capture 3D data in real-world settings. Specifically, 3D facial data is crucial for anthropometric studies. However, in uncontrolled environments, the collected data may include extraneous information, such as clothing or anything below the neck. Additionally, faces may not be globally registered and could even be considered cross-dataset, necessitating a further global registration and alignment step using standard methods like ICP. Dataset that will be used are YORK and Facescape. All faces are labeled with 68 facial landmarks (with different semantics in each dataset).

Final Objective of the Study

The aim of this project is to implement different techniques to preprocess the data, preparing it for subsequent learning tasks, such as 3D facial landmark detection.

Methodology:

- 3D mesh cleaning (hole filling, statistical outliers removal, etc.)
- 3D face segmentation (via classical methods or learned methods)
- Global registration across two datasets (Facescape and Headspace datasets)
- Produce metrics in order to visualize the method effectiveness

5) Landmark Detection Refinement (GIA' ASSEGNATO)

Study Description

Our most recent work focused on 3D Facial Landmark Localization using a GNN-based method. This network operates on a specific representation of a resampled 3D face point cloud as a graph, where mesh vertices are considered nodes, and the adjacency matrix is constructed using k-NN. For each node, the network produces a vector of L features (where L is the number of landmarks), representing the probability values for a vertex to be the i-th landmark (with i=1,...,L). The network's output is essentially a heatmap, where each point has L probability values. It is quite evident that selecting a point on the resampled version of the mesh leads to poor performance. Although the resampled points still lie on the original surface, they do not belong to the set of points of the original mesh where the true landmark can be found.

To locate a specific landmark, one approach is to select the point with the highest probability value from the corresponding heatmap. However, this method is misleading because the heatmap points come from the resampled version of the mesh, not the original one. An alternative approach is to find the closest point on the original surface. Nevertheless, we found that solely relying on the maximum value from the heatmap leads to poor performance.

Final Objective of the Study

The aim of this project is to produce different techniques to infer true landmark position starting from a predicted or GT 3D heatmap and producing metrics to compare different methods effectiveness

Methodology:

DISCLAIMER: ALL GT HEATMAPS HAVE BEEN PRODUCED FOLLOWING THIS METHOD

- For each face you start from a predicted or GT heatmap (a tensor of N_POINTS x 3 x L) the true mesh and the true landmark positions (x,y,z)
- Produce at least 3 techniques to infer true landmark positions
- EXTRA (learned method)

Specifically, given the resampled vertices V of a 3D face point cloud, and the corresponding landmarks L, for each vertex $v \in V$, we compute the Euclidean distances to every landmark $l \in L$, producing a distance matrix Δ with dimensions $\mathbb{R}^{|V|\times|L|}$. This matrix is then transformed using a Gaussian function to encode the distances into a normalized probability matrix, represented as the 3D heatmap:

$$H = \exp\left(-\frac{\Delta^2}{2\sigma^2}\right) \in \mathbb{R}^{|V|\times|L|}, \tag{2}$$

where σ represents a hyperparameter that controls the spread of the heatmap. Essentially, the 3D heatmap indicates the probability of each landmark occurring at a specific location as shown in Figure 2 for one example case.

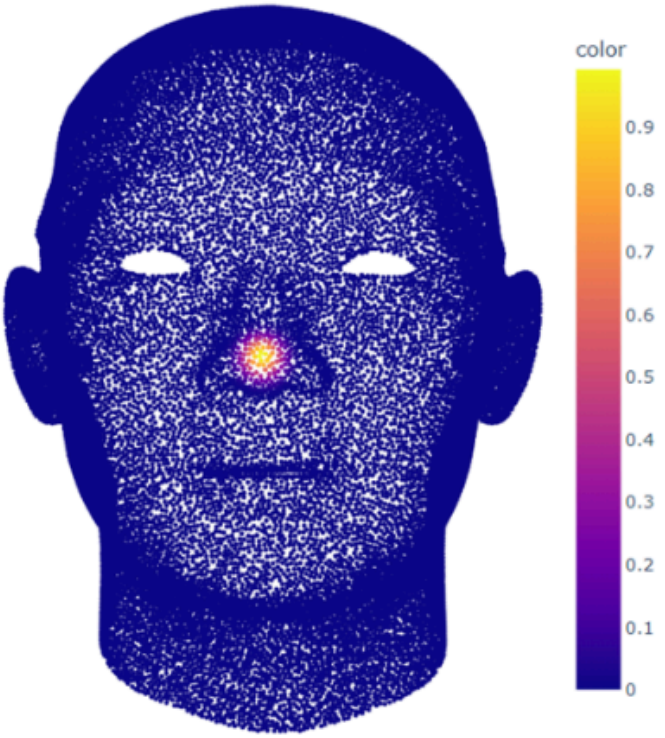
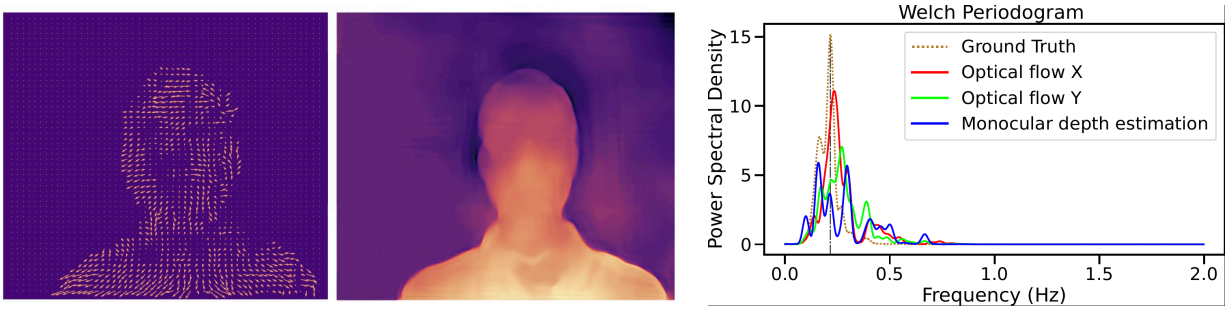


Fig. 2. An example of a 3D heatmap corresponding to the nose landmark. It is obtained applying Eq. 2 with $\sigma = 0.03$.

6) Monocular Scene Flow Estimation

Monocular scene flow estimation deals with obtaining 3D structure and 3D motion from two temporally consecutive RGB images. The project consists of applying one (or more) Monocular Scene Flow Estimators on videos showing people and using the extracted estimates to approximate respiration induced movements.



Here's (a non comprehensive list) of approaches estimating Scene Flow Monocularly:

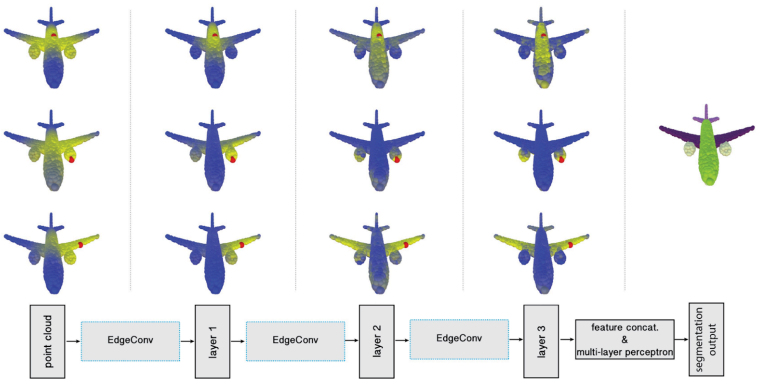
- https://github.com/visinf/self-mono-sf (Code), https://openaccess.thecvf.com/content_CVPR_2020/papers/Hur_Self-Supervised_Monocular_Scene_Flow_Estimation_CVPR_2020_paper.pdf (Paper)
- https://github.com/BlarkLee/MonoPLFlowNet (Code), https://arxiv.org/pdf/2111.12325 (Paper)

7) E(n)-Equivariant Dynamic Graph CNN for point cloud segmentation

(Technically Sound)

One of the research objective in our lab is 3D point cloud understanding, particularly focusing on 3D segmentation**.** A widely used architecture for this purpose is DGCNN [1][1-Code] (*Dynamic Graph Convolutional Neural Network*). Unlike traditional convolutional neural networks that operate on a regular grid, DGCNN can effectively process unstructured point cloud data. DGCNN excels in capturing local geometric structures by dynamically constructing graphs based on the k-nearest neighbors in the feature space, reaching state-of-the-art performances in many tasks, including 3D segmentation. However, a significant drawback of DGCNN is its lack of invariance to translation, rotation, and reflection which is pillar to many real-life applications. To overcome this limitation, **this project propose to incorporate E(n)-equivariant message passing [2] into the DGCNN architecture** (which is possible due to the invariance of scalars values, which in this case would correspond to the probability value that a generic point of the original point cloud belongs to a particular segment rather than another, see Figures.)

The dataset to use is ShapeNet [3].



[1] https://arxiv.org/pdf/1801.07829

[1-Code] https://github.com/WangYueFt/dgcnn

[2] https://arxiv.org/pdf/2102.09844

[3] https://arxiv.org/abs/1512.03012