

PAC. Pràctica 2.
J. de Curtò i DíAz & I. de Zarzà i Cubero.
c@decurto.be z@dezarza.be

Tipologia i cicle de vida de les dades. Màster de Ciència de Dades.

Alumnes/as:

J. de Curtò i DíAz. decurto@uoc.edu

I. de Zarzà i Cubero. dezarza@uoc.edu

Pràctica 2. Neteja i anàlisi de dades.

Extensió del dataset CyZ, generat a la Pràctica 1.

CyZ: MARS Space Exploration Dataset.

<https://github.com/decurtoidiaz/cyz>

DrCyZ: Techniques for analyzing and extracting useful information from CyZ.

<https://github.com/decurtoidiaz/drcyz>

github.com/decurtoidiaz/cyz

readme.md

CyZ

DOI [10.5281/zenodo.5655473](https://doi.org/10.5281/zenodo.5655473)

CyZ: MARS Space Exploration Dataset.

Images from NASA missions of the celestial body.

Images from the Mars Perseverance Rover

NASA

MARS 2020 MISSION PERSEVERANCE ROVER

Multimedia - Raw Images

View More Images

Raw Images

268 new images

161,169 total images

Showing 401-500 of 66,622 results

<

1

2

3

4

5

6

7

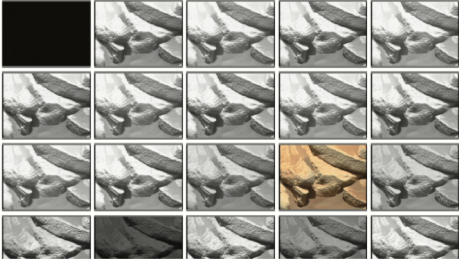
8

9

10

of 667

>



Filters

Sort by Newest to Oldest

Dates Latest Images

From 1997-01-01 To 2021-01-01

Image Type

☐ Raw

☐ Color-Processed

Engineering Cameras

☐ Navigation Camera - Left

f

t

o

o

J. de Curtò i DíAz c@decurto.be

I. de Zarzà i Cubero z@dezarza.be

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

El dataset [1] conté imatges de les missions Perseverance i Curiosity de la NASA [2,3,4,5,6] i està inspirat en datasets per dur a terme tasques de visió per computador en aplicacions espacials [7,8]. Va ser introduït prèviament a la pràctica 1 de web scraping d'aquesta assignatura. Aquest dataset presenta imatges que poden ésser molt útils per fer recerca en visió per computador i el nostre objectiu és netejar les dades, fer una anàlisi en profunditat d'aquestes i estendre la informació proporcionada a la pràctica anterior.

2. Integració i selecció de les dades d'interès a analitzar.

El primer pas a dur a terme és una anàlisi a través del k-means clustering per visualitzar les dades de manera que sigui possible dur a terme la seva neteja. També ens permetrà identificar els grups d'imatges que s'assemblen més entre ells i per tant, les càmeres que ens poden ser més útils a seleccionar donada una tasca de recerca posterior. Un cop seleccionades les dades que utilitzarem per fer les posteriors anàlisis, es pot optar primer per estudiar les dades de manera estadística i després per utilitzar mètodes més sofisticats d'aprenentatge automàtic i profund per una tasca concreta (classificació, segmentació, generació d'imatges sintètiques...).

3. Neteja de les dades.
 1. 3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?
 2. 3.2. Identificació i tractament de valors extrems.

Apliquem K-means clustering i visualitzem la informació per detectar valors a excloure. També fem un estudi per buscar imatges que no contenen informació o que estan borroses i no aporten dades significatives. A més a més, també plantejem l'aplicació de la tècnica Principal Components Analysis (PCA) [9] per entendre e interpretar la informació que aporta cadascuna de les càmeres.

4. Anàlisi de les dades.
 1. 4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).
 2. 4.2. Comprovació de la normalitat i homogeneïtat de la variància.
 3. 4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

Apliquem la tècnica T-distributed Stochastic Neighbor Embedding (T-SNE) [10,11,12,13] per fer una visualització de les dades; es tracta d'una tècnica no-lineal de reducció de la dimensionalitat. Primer es construeix una distribució de probabilitats sobre parells d'imatges de tal manera que s'assigna probabilitat més alta a les instàncies semblants mentre que s'assigna baixa probabilitat a les mostres que no s'assemblen. Després es projecten aquests punts en un mapa de baixa dimensionalitat i es minimitza la KL divergence [14] entre les dues distribucions. Aquesta tècnica ha estat molt emprada per visualitzar imatges en el context de datasets per visió per computador i moltes altres aplicacions (genòmica, nlp, medicina,...).

De manera paral·lela, també adjuntem una anàlisi de la normalitat i la variància del conjunt de dades seleccionades. Definirem un seguit de variables, com per exemple tipus de càmera i intensitat de la imatge per aplicar diferents proves estadístiques. (a afegir a la versió final)

A més a més, utilitzarem una tècnica d'aprenentatge no supervisat, Generative Adversarial Networks (GAN) [15] per generar a partir del dataset proporcionat noves imatges sintètiques i veure si la xarxa és capaç d'extreure la informació més significativa. Aquest procés ens permet entendre millor l'estructura de les mostres i la capacitat per generar noves dades donat un subconjunt de dades reals.

Per últim, usarem una tècnica de segmentació semàntica [16] entrenada en un dataset terrestre amb imatges de paisatges [17] per proporcionar màscares semàntiques aproximades de les imatges. Això ens permetrà entendre el contingut visual d'aquestes i detectar imatges que no aporten al conjunt i aquelles que són més significatives per una tasca concreta.

5. Representació dels resultats a partir de taules i gràfiques.

Es realitzarà la presentació de les anàlisis anteriors de manera visual, mitjançant també exemples de quan funcionen bé els algorismes i quan no (en el cas de segmentació semàntica i generació sintètica d'imatges).

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions?

Els resultats permeten respondre al problema?

Durem a terme un anàlisi dels resultats obtinguts.

7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

El codi en Python es pot trobar al repositori:

DrCyZ: Techniques for analyzing and extracting useful information from CyZ.
<https://github.com/decurtoidiaz/drcyz>

Referències:

- [1] De Curtò and De Zarzà. CyZ: MARS Space Exploration Dataset. Zenodo. 2021.
<https://doi.org/10.5281/zenodo.5655473>
- [2] Spirit: <https://mars.nasa.gov/mer/gallery/all/spirit.html>
- [3] Opportunity: <https://mars.nasa.gov/mer/gallery/all/opportunity.html>
- [4] Curiosity: <https://mars.nasa.gov/msl/multimedia/raw-images/>
- [5] Perseverance: <https://mars.nasa.gov/mars2020/multimedia/raw-images/>
- [6] Maki et al. 2020. The Mars 2020 Engineering Cameras and Microphone on the Perseverance Rover: A Next-generation Imaging System for Mars Exploration.
<https://link.springer.com/article/10.1007/s11214-020-00765-9>
- [7] Lamarre et al. 2020. The Canadian Planetary Emulation Terrain Energy-Aware Rover Navigation Dataset.
<https://starslab.ca/enav-planetary-dataset/>
- [8] ESA Robotics Dataset. 2015. Katwijk Beach Planetary Rover Dataset.
<https://robotics.estec.esa.int/datasets/katwijk-beach-11-2015/>

[9] Shlens. A Tutorial on Principal Component Analysis. 2005.

<https://www.cs.cmu.edu/~elaw/papers/pca.pdf>

[10] Van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research 15(Oct):3221-3245, 2014.

[11] Van der Maaten and G.E. Hinton. Visualizing Non-Metric Similarities in Multiple Maps. Machine Learning 87(1):33-55, 2012.

[12] Van der Maaten. Learning a Parametric Embedding by Preserving Local Structure. In Proceedings of the Twelfth International Conference on Artificial Intelligence & Statistics (AI-STATS), JMLR W&CP 5:384-391, 2009.

[13] Van der Maaten and Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

[14] https://en.wikipedia.org/wiki/Kullback-Leibler_divergence

[15] Goodfellow et al. Generative Adversarial Networks. NIPS. 2014.

[16] He et al. Mask R-cnn. ICCV. 2017.

[17] Zhou et al. Semantic Understanding of Scenes through the ADE20K Dataset. IJCV. 2016.

| Contribucions | Signatura |
|---------------------------|-----------|
| Investigació prèvia | JDC, IDZ |
| Redacció de les respostes | JDC, IDZ |
| Desenvolupament del codi | JDC, IDZ |