Analysis and interpretation of the LFW dataset and investigating qualitative methods using inferential statistics - Research Methods Assignment 2 Resit

**Task 1 Analyse a non-trivial research dataset**

The chosen datasets to be analysed were face recognition dataset containing photos of people that were used to create a machine learning model that could then identify people with a high degree of accuracy. In order to apply inferential statistics onto this type of data, more information was necessary. Information such as the Percentage of each ethnicity present in the dataset. As this information was not publicly available anywhere, an ethnicity detection algorithm was used to iterate through the dataset and classify each person as their most likely ethnicity. The Deepface python library which boasts an accuracy of 97.44% was used for this. The below table illustrates the percentage of each ethnicity present in the Labelled Faces in the wild dataset.

Table

Description automatically generated

Fig. This results of the analysis proves the accuracy of the ethnicity detection algorithm to be within the margins of the states accuracy level with an accuracy of 97% (57+7+10+10+12+1) in comparison to the stated 97.44%. However, these results are also proof that the dataset is skewed as 57% of the sample size is of a white ethnic background. The “race unrecognised” section does NOT represent wrongly classified ethnicities, instead it represents people who’s ethnicities could not be detected due to blurry or inadequate images.

After acquiring the percentages of each Ethnicity in the dataset, 5 samples of 20 people from each ethnicity was chosen to test the face recognition accuracy of the dataset by comparing, within each ethnic group, a person from the 20 at random with the other 19 participants and accessing the M.A.D (Mean Average Distance), a unit of measurement that measured how similar two faces being compared are. A higher Distance value would mean the two faces being compared are less similar and a lower distance value would mean the faces being compared are more similar. Therefore, the higher the mean Average, the better the algorithm is at correctly identifying a person from an ethnic group.

When comparing the results between black and white people, the results illustrated that the algorithm was better at differentiating between white people by at least 9%. Given that only a sample of 20 was tested and there are only 425 black people in comparison to a staggering 3277 white people in the dataset, it is evident that the algorithm would person even better if a larger sample was tested as the Face recognition model generated with this dataset would have significantly more samples of white people to be trained on, making it significantly better at recognising white people. Below are the T and Z test results of the comparison.

|  |  |  |
| --- | --- | --- |
| t-Test: Two-Sample Assuming Unequal Variances | Black Ethnicity | White Ethnicity |
| Mean | 0.666842 | 0.751053 |
| Variance | 0.013001 | 0.013677 |
| Observations | 19 | 19 |
| Hypothesized Mean Difference | 0 |  |
| df | 36 |  |
| t Stat | -2.24736 |  |
| P(T<=t) one-tail | 0.01542 |  |
| t Critical one-tail | 1.688298 |  |
| P(T<=t) two-tail | 0.030839 |  |
| t Critical two-tail | 2.028094 |  |

**Task 2: Investigating existing research methods and background literature for acquiring and analysing similar datasets.**

Collecting face data for face recognition systems is sometimes difficult due to security and privacy-related concerns of people. In addition, choosing the right method is a crucial step in simplifying the facial data collection process, therefore the data should meet certain criteria and certain factors need to be outlined to ensure the right data is collected. These factors constrain the following:

* What data collection means for facial recognition systems
* The lighting conditions the images were taken in
* The size of the images
* If there are multiple faces in the image

Face recognition is a type of machine learning and a branch of image recognition that falls under image data collection. It involves gathering face images to train and improve Face Recognition Systems.

The process begins by taking annotated face images of different people and feeding them into a machine learning model which uses them to learn how to scan, identify and process facial features using techniques from one or multiple face detection methods available such as the HAAR Cascade method with Viola Jones Potential, YOLO or SVM to name a few.

Some of the most common methods of collecting data for facial recognition systems include but are not limited to:

**Pre-packaged/Public face image datasets**

Public datasets are created by a third party and can either be open source or bought. Labelled Faces and CelebA are examples of free facial datasets.

The advantages of this method are that it is easy to access as the datasets are readily available online and can be downloaded quickly. Furthermore, it is cheaper than in-house data collection if not free to use.

The disadvantages of this method are that there is a lack of customization. As the data is already collected, there is little that can be done to categorise the sample to make it specific for a unique project’s requirements. If the dataset does not have images with a certain type of obstruction on the face such as masks or glasses, then additional data needs to be added and this can add pre-processing costs to a budget.

In addition to a lack of customization, the average quality Level of the images can be low, and this can affect the performance of the generated model. This may be because some public datasets to not go through rigorous quality checks and are collected by the public therefore the credibility of the dataset can also be questionable.

Another method of collecting face data is through crowdsourcing

**Face image data collection through crowdsourcing**

Crowdsourcing is appealing in that it combines working with the public to gather fresh face image datasets with an in-house infrastructure that allows the submission of data in exchange for compensation. This may be through an online platform that allows crowds to register, get data collection projects and then submit them where appropriate such as a contest or competition.

The advantages of crowdsourcing are that is scalable and customisable, this can be done by specifying your requirements to the service provider.

It is cheaper than in-house since the crowd uses their own devices.

The data collected is often more diverse as more diverse people are involved in the data collection process since a larger number of contributors can be reached worldwide as opposed to only utilising a local sample size that can be very biased depending on location.

Lastly, there are no copyright issues since when the crowd registers they are required to agree to terms and conditions that transfer over rights of the image data to the researchers.

Unfortunately, no methods are free from disadvantages. The demerits of crowdsourcing are that if constraints are not set at the start then there can be quality issues since the devices that are used for data collection will vary, meaning that data will come in varying sizes, of varying quality and quantity and with expectations of compensation all the same. The problem with adding constraints to rectify this would be that the sample size would then decrease to only those of people who meet the criteria. However, this could be an effective strategy for ensuring that all data collected will be compatible with the machine learning model and algorithms that will be applied to the data.

Lastly, as crowdsourcing is typically done through an online platform such as a mobile application, developing or purchasing such software cab add extra time and costs to the process.

Now having looked at two methods which rely on human participants, it is time to consider alternatives when this falls short or other methods are more appropriate. Automated face image data collection is one such method.

**Automated face image data collection**

It is possible to automate image data for face recognition systems by integrating machine learning into the process. This can be done through web scraping or crawling techniques, where the data us scraped from specified or unspecified online sources.

The advantages of this are that no human input is required, requirements can be easily specified and are sure to be followed as this would be free from human errors. Lastly, it is cheaper since only the machine learning model and system it is running on is required, no additional equipment or recruitment of contributors is required.

The disadvantages of automation through machine learning is that, there will be additional pre-processing costs as the scraped/crawled data will need to be cleaned and processed.

Moreover, there will be a limitation from anti-scraping techniques as online sources often use various anti-scraping techniques to block web scraping.

The last method being considered is arguably the most reliable but also most expensive method and that will be the In-house face image data collection method.

**In-house face image data collection**

The inhouse data collection method involves creating a separate image data collection project to develop a facial recognition system and this will require a team to purchase cameras and lighting equipment as well as hire contributors to take their images.

The advantages of this are that there will be a higher data protection factor which is a bonus for projects that are confidential in nature.

In house also provides more control as the team can control every aspect of the process such as choosing which cameras to use, the contributors they hire and the setting, lighting, background etc.

Lastly, In house productions comes with complete data ownership as since an image of someone’s face is their biometric information, getting ownership rights is important before using it and inhouse image data collection allows complete ownership of the data by the company and eliminates the risk of future data-related lawsuits as the contributors sign an agreement with the company during recruitment.

The only disadvantages of in-house production of data are that, this can cause a lack of diversity as the inhouse method limits the level of diversity in the dataset since a smaller sample size will be hired in comparison to crowdsourcing.

In addition to this, in-house is the most expensive method since the company has to cover all the expenses by itself.

To finish off, In-house is also time-consuming in comparison to using pre-packaged data or crowdsourcing since the company has to prepare for and gather the data itself as well as processing it all.

References:

S. Javaid, Top 4 Facial Recognition Data Collection Methods in 2023, Available at: https://research.aimultiple.com/facial-recognition-data-collection/ Accessed [12/1/2023]