Selected Domain

In several countries it is common for people to harvest mushrooms in the wild for consumption. A common danger with this activity is people are not aware about the potential toxicity of the mushrooms they pick. In Spain alone there are 400 cases of severe intoxication a year[1]. We propose to use the power of Content-Based Image Retrieval (CBIR) to build an application that can identify the species of a mushroom based on a provided image.

The reason CBIR is the correct tool for this issue is people will likely not know the name of the species they are looking at and it is risky to allow them to use key word text to look up the species. It is far more appropriate to use the distinct features of each mushroom species, such as shape, colour and texture, via image indexing as the searching mechanism. Previous work shows that this is a proper domain for application[2].

Data

To retrieve the dataset, we scrapped the page http://www.mushroom.world/, which provides labeled data for hundreds of different mushroom species. We reused an open source implementation from a group of Helsinki University students[3]. This provided us with a dataset of 626 labeled images. With the set consisting of at least 2 images of each species of mushroom it contains. Due to memory constraints the final results were generated using 26 images made up of 2 images of 13 different species.

LIRE Extension

For the implementation of our LIRE extension, we will use two steps in our extractor, using SURF[4] for keypoint detection and a Color Histogram to describe the image features (which we will implement).

- Speeded-Up Robust Features[5] (SURF) is an algorithm that approximates the Laplacian of Gaussian acting as a blob detector which detects blobs in various sizes. The reason is that it is a very costly operation, and Box Filters are used in order to filter the original image and reduce the workload of the operation. This method was chosen to help identify the distinct shape of mushrooms in our images.
- Color Histogram is a representation of the distribution of colors in an image. For digital images, a color histogram represents the number of pixels that have colors in each of a fixed list of color ranges, that span the image's color space. This method was selected as generally each species of mushroom has a unique pallet of colours.

Once an image has been reduced to a colour histogram, a search can be performed by comparing it to other image histograms. After reviewing some literature on the topic of Color Histogram Based Image Retrieval[6], we discovered multiple algorithms for searching images based on the similarity between two images, such as Euclidean Distance, Histogram Intersection or Earth Mover's distance.

[1] https://www.heraldo.es/noticias/salud/2018/10/08/cada-ano-producen-400-casos-graves-intoxicaciones-por-setas-1270707-2261131.html

[2] https://pdfs.semanticscholar.org/8efa/ea4085e64785143e21f1797e9c2c95c8f2f7.pdf

[3] https://github.com/TuomoNieminen/deep-shrooms

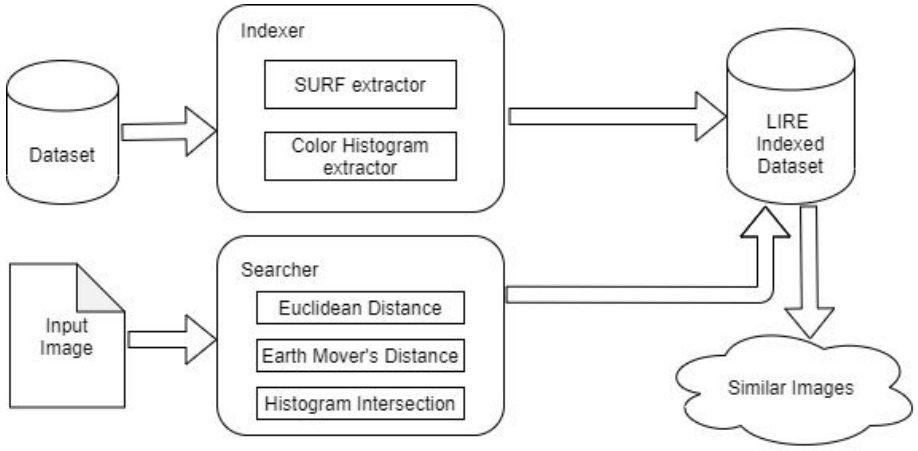
Euclidean Distance	$D_{L2} = \sqrt{\sum_i (h_1(i) - h_2(i))^2}$	Distance between two points in space corresponds to the length of a straight line drawn between them.
Earth Mover's	$D_{EM} = \frac{min_{f_{ij}} \sum_{i,j} f_{ij} A_{ij}}{sum_{i,j} f_{ij}}$	Distance between two probability distributions over a region. Minimum cost of turning one histogram into another.
Histogram Intersection	$D_\cap = 1 - \frac{\sum_i (min(h_1(i), h_2(i))}{min(\lvert h_1(i) \rvert, \lvert h_2(i) \rvert)}$	Similarity of two discrete probability distributions.

Design and Implementation

The final implementation and all the code can be found in the following Github repository:

https://github.com/rguitar96/lire-ir

For the final design, we implemented a LIRE application that first indexes the dataset using two feature extractors: SURF and a custom Histogram Extractor, implemented in the file ColorHistogramExtractor_global. After storing the data, the user can search with input images of mushrooms, receiving similar images as output:



After experimenting with different colour spaces it was deduced that RGB was the most robust method for describing image features. Then by experimenting with the different search method euclidean distance gave acceptable results when matching.

Results

For our results, we defined a metric classifying the images in three categories:

- **Same species** is in the **top 1**: the program recognized perfectly the species, returning a mushroom of the same species as the first result. Test set result: 14/26 (**54%**). For example, the algorithm very often detects the species **Laccaria Laccata**.
- **Same species** is in the **top 10**: the program returned a mushroom of the same species within the first ten output images. Test set result: 9/26 (**35%**). For example, the species **Cantharellus Cibarius** falls in this category.
- There is **no same species** in the top 10: the program did not return the proper species within the first 10 results. Test set result: 3/26 (**11%**). The algorithm did not correctly detect the species **Gomphidius Glutinosus**.

Results for each image can be seen in the folder "SimpleApplication-1.0b4/data/searchResults"

[4] https://www.researchgate.net/figure/Laplacian-of-Gaussian-approximation-with-box-filters-SURF-keeps-the-size-of-the-input_fig3_324782189

[5] https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_feature2d/py_surf_intro/py_surf_intro.html

[6] https://www.researchgate.net/publication/224503719_A_Study_of_Color_Histogram_Based_Image_Retrieval