Visualization-Guided Classification of Carbonized Seeds from Early Human Civilizations

Andreas Gogel
Visual Computing

**Motivation**
Approximately about 10,000 years ago, humans began the transition from a hunter and gatherer to a settlement society. This process is called the Neolithic Revolution and one of the hot spots of this shifting lies in the Fertile Crescent located in the Near East. Researchers in the field of Botanical Archeology are trying to understand which plant species were selected and how they changed gradually. Most of the remaining seeds from these very early times are carbonized. This way, they could be preserved. However, in this condition, it is hard to identify the species of the seeds. During the burning process, the seeds are deformed and loose surface details. Normally, an expert is consulted to classify them. Since there are only a few experts in this field, an automatic approach is requested.

**Data Challenges**
The data set (over 400 GB) consists 1043 labeled seeds of 22 different species (scanning process ongoing). The distribution of the samples is very unbalanced (some species contain over 200 seeds, others only a single seed). They can be divided into archeological seeds, fresh recent seeds and in the laboratory carbonized seeds to increase the ground truth. Since some species are already extinct, at least a suggestion of the biological genus should be given.

**Results**

**Tool Widget**
- Switch between single and multiselection (scatter plot)
- Delete selected samples (scatterplot) e.g., outliers or incorrectly segmented seeds
- Accuracy of the classifier with color coded background

**Species Overview Widget**
Shows the median seed of every species (fresh and carbonized) according to the selected features with color coded frames. The median is selected to reduce the influence of outliers. The seeds are all equally rotated. Rotation (linked between all widgets) and zooming is possible. If there are no fresh or carbonized samples available, the widget stays black. Selections of species affect also the scatter plot. Different rendering methods are possible (Task 2).

**Scatter Plot Widget**
2D scatter plot of the selected features with additional visualizations (Task 2). Median seeds are marked with a colored ring. The suggestions of the classifier (soft voting, containing a decision tree and a random forest) are shown in the right upper box with the probabilities greater than zero of all classes (Task 3).

**Example classifications**
- Lens_tenuiflorus: 10.66%
- Vicia_Sativa: 88.75%
- Lathyrus_inclusus: 1.25%
- Lens_ervoides: 10.25%
- Lens_orientalis: 45.87%
- At least biological genus classified (Lens): 86.6%

**Conclusion**
The features, the trained model and separated seeds can be exported (Task 1) and the classifier has an accuracy of approximately 85% (90% on genus level). The domain expert can check and manually correct the suggestions of the classifier and is used for an international project.