MIDI 2025 Conference

Prints in the Magnetic Dust: Robust Similarity Search in Legacy Media Images Using Checksum Count Vectors

Abstract

Digitizing magnetic media containing computer data is only the first step towards the preservation of early home computing era artifacts. The audio tape images must be decoded, verified, repaired if necessary, tested, and documented. If parts of this process could be effectively automated, volunteers could focus on contributing contextual and historical knowledge rather than struggling with technical tools. We therefore propose a feature representation based on Checksum Count Vectors and evaluate its applicability to detecting duplicates and variants of recordings within a large data store.

The approach was tested on a collection of decoded tape images (n=4902), achieving 58% accuracy in detecting variants and 97% accuracy in identifying alternative copies, for damaged recordings with up to 75% of records missing.

These results represent an important step towards fully automated pipelines for restoration, de-duplication, and semantic integration of historical digital artifacts through sequence matching, automatic repair and knowledge discovery.