

# Implementing the Context Tree Weighting Method for Content Recognition

Zaher Dawy, Joachim Hagenauer, and Andreas Hoffmann

Institute for Communications Engineering (LNT), Munich University of Technology (TUM)  
Arcisstr. 21, 80290 Munich, Germany, Email: Zaher.Dawy@ei.tum.de

The context tree weighting method (CTW) is a statistics-based universal data compression algorithm that is capable of achieving superior performance compared to Lempel-Ziv based algorithms [1], [2]. Motivated by this fact, we investigate the usability of CTW for applications involving content recognition. Recently, various authors have explored the application of other data compression algorithms for content recognition, e.g. see [3], [4], [5]. Given a test file that needs to be classified among a set of several reference files that represent different classes, the reference file which leads to the best compression of the test file when both files are appended is selected as the most probable match. Moreover, we modify CTW for content recognition purposes by introducing the concept of context tree *freezing* after the reference sequence is encoded to avoid learning the memory structure of the appended test sequence. Results show that CTW with the proposed *freezing* technique achieves a clearly superior performance compared to a wide range of other compression algorithms for content recognition problems such as language recognition, authorship attribution, and DNA data classification. For more details, the reader is referred to the full paper version available at [6].

## REFERENCES

- [1] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context tree weighting method: Basic properties," *IEEE Trans. Info. Theory*, vol. 41, pp. 653–664, May 1995.
- [2] K. Sadakane, T. Okazaki, and H. Imai, "Implementing the context tree weighting method for text compression," in *IEEE Data Compression Conference*, (Snowbird, Utah, USA), March 2000.
- [3] W. J. Teahan, S. Inglis, J. G. Cleary, and G. Holmes, "Correcting english text using PPM models," in *IEEE Data Compression Conference*, (Snowbird, Utah, USA), March 1998.
- [4] D. Benedetto, E. Caglioti, and V. Loreto, "Language trees and zipping," *Physical Review Letters*, vol. 88, January 2002.
- [5] A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, and A. Vulpiani, "Data compression and learning in time sequence analysis," *Physica D*, vol. 180, pp. 92–107, January 2003.
- [6] Institute for Communications Engineering, Munich University of Technology, <http://www.lnt.ei.tum.de>.