

Off-line Compression by Extensible Motifs

Alberto Apostolico*
Purdue Univ. & Univ. of Padova

Matteo Comin†
Univ. of Padova

Laxmi Parida‡
IBM T. J. Watson Center

We present lossy off-line data compression techniques by textual substitution in which the patterns used in compression are chosen among the *extensible motifs* that are found to recur in the textstring with a minimum pre-specified frequency. A motif is to be interpreted here as a sequence of intermixed solid and don't care characters that obeys, in addition, some conditions of saturations: most notably, it must be not possible to eliminate some don't cares in the pattern without having to forfeit some of its occurrences. Motif discovery and motif-driven parses of various kinds have been previously introduced and used in [1, 2]. Whereas the motifs considered in those studies are "rigid", here we assume that each sequence of gaps present in a motif comes endowed with some individually prescribed degree of elasticity, whereby a same pattern may be stretched to fit segments of the source that match at all the solid characters but are otherwise of different lengths. This is expected to save on the size of the codebook, and hence to improve compression.

Traditionally, the design of codebooks used in compression proceeds from specifications that are either statistical or syntactic. Of course, these two aspects blend in the final code. With Huffman codes, for instance, once the characters are statistically ranked a code with certain syntactic characteristics, notably, obeying the prefix property, is built. Likewise, once the codebook of an error correcting code is designed, the statistics of the source is taken into account for encoding. However, these two stages are, as a rule, carried out somewhat independently. By contrast, the notion of a motif that we adopt tightly combines the structure of the motif pattern, as described by its syntactic specification, with the statistical measure of its occurrence count. This supports a notion of saturation that mandates that motifs that could be made more specific without altering their set of occurrences do not bear interest and may be discarded.

The transition from rigid to extensible motifs requires the following ingredients, the predisposition and orchestration of which form the contribution of the present paper: A algorithm for the extraction of flexible motifs; A criterion for choosing and encoding the motifs to be used in compression; A new suite of software programs implementing the whole.

We are encouraged by the fact that in images a tremendous amount of compression is attained, albeit with a large loss such as 40% or so, yet simple predictors in the form of linear interpolation restore more than 95% of the original.

References

- [1] A. Apostolico, M. Comin and L. Parida, "Motifs in Ziv-Lempel-Welch Clef" Proceedings of *IEEE DCC Data Compression Conference*, pp. 72–81 Computer Society Press, (2004).
- [2] A. Apostolico and L. Parida, "Compression and the Wheel of Fortune", *Proceedings of DCC 2003*, IEEE Computer Society Press, 143–152 (2003).

*Department of Computer Sciences, Purdue University, Computer Sciences Building, West Lafayette, IN 47907, USA and Dipartimento di Ingegneria dell'Informazione, Università di Padova, Padova, Italy. Work Supported in part by the Italian Ministry of University and Research under the National Projects FIRB RBNE01KNFP, and PRIN "Combinatorial and Algorithmic Methods for Pattern Discovery in Biosequences", and by the Research Program of the University of Padova. axa@dei.unipd.it

†Dipartimento di Ingegneria dell'Informazione, Università di Padova, Padova, Italy. Work done during internship at IBM Thomas J. Watson Research Center. ciompin@dei.unipd.it

‡IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA. parida@us.ibm.com