

# Improving Run Length Encoding (RLE) on bit level by preprocessing

## Bachelorarbeit

zur Erlangung des akademischen Grades  
Bachelor of Science (B.Sc.)

Universität Trier  
FB IV - Informatikwissenschaften  
Lehrstuhl für Informatik I

Gutachter:	Prof. Dr. Ingo J. Timm xxxxxxxxxx
Betreuer:	xxxxxxxxxx

Vorgelegt am xx.xx.xxxx von:

Sven Fiergolla  
Am Deimelberg 30  
54295 Trier  
sven.fiergolla@gmail.com  
Matr.-Nr. 1252732

# Zusammenfassung

Hier steht eine Kurzzusammenfassung (Abstract) der Arbeit. Stellen Sie kurz und präzise Ziel und Gegenstand der Arbeit, die angewendeten Methoden, sowie die Ergebnisse der Arbeit dar. Halten Sie dabei die ersten Punkten eher kurz und fokussieren Sie die Ergebnisse. Bewerten Sie auch die Ergebnissen und ordnen Sie diese in den Kontext ein.

Die Kurzzusammenfassung sollte maximal 1 Seite lang sein.

# Inhaltsverzeichnis

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Problem statement . . . . .	1
1.3	Main Objective . . . . .	2
1.4	Gliederung/Aufbau der Arbeit . . . . .	2
<b>2</b>	<b>Basic principles of compression</b>	<b>3</b>
2.1	Compression and Encoding fundamentals . . . . .	3
2.2	Run Length Encoding . . . . .	3
2.3	State of the Art . . . . .	3
<b>3</b>	<b>Analyse</b>	<b>4</b>
3.1	Anforderungen . . . . .	4
3.2	Weiterer Abschnitt . . . . .	4
3.3	Zusammenfassung . . . . .	5
<b>4</b>	<b>Entwurf / Konzeption</b>	<b>6</b>
4.1	Abschnitt 1 . . . . .	6
4.2	Abschnitt 2 . . . . .	6
4.3	Zusammenfassung . . . . .	7
<b>5</b>	<b>Implementierung</b>	<b>8</b>
5.1	Abschnitt 1 . . . . .	8
5.2	Abschnitt 2 . . . . .	8
<b>6</b>	<b>Evaluation</b>	<b>9</b>
6.1	Abschnitt 1 . . . . .	9
6.2	Abschnitt 2 . . . . .	9
6.3	Zusammenfassung . . . . .	9
<b>7</b>	<b>Diskussion und Ausblick</b>	<b>10</b>
	<b>Literaturverzeichnis</b>	<b>11</b>

# Abbildungsverzeichnis

3.1	Logo der Universität Trier. . . . .	4
-----	-------------------------------------	---

# Tabellenverzeichnis

3.1 Tabelle mit Werten. . . . . 4

# 1. Introduction

Die Einleitung besteht aus der Motivation, der Problemstellung, der Zielsetzung und einem ersten Überblick über den Aufbau der Arbeit.

TODO:

- explain compression ratio
- define unit of compression to quantify question and results

## 1.1 Motivation

In the last decades, digital data transfer became available everywhere and to everyone. This rise of digital data urges the need for data compression techniques or improvements on existing ones. Run-length encoding (abbreviated as rle) is a simple coding scheme that performs lossless data compression. rle compression simply represents the consecutive, identical symbols of a string with a run, usually denoted by  $\sigma^i$ , where  $\sigma$  is an alphabet symbol and  $i$  is its number of repetitions. To give an example, the string aaaabbaaabbba can be compressed into rle format as  $a^4b^2a^3b^4a^1$ . Its simplicity and efficiency make run-length encoding still used in several areas like fax transmission, where rle compression is combined with other techniques into Modified Huffman Coding [?]. Most fax documents are typically simple texts on a white background, rle compression is particularly suitable for fax and often achieves good compression ratios. Another application of rle is optical character recognition, in which the inputs are usually images of large scales of identically valued pixels. Other applications appear in bioinformatics, where rle compression is employed to speed up the comparison of two biological sequences.

## 1.2 Problem statement

Some strings like aaaabbbb archive a very good compression rate because the string only has two different characters and they repeat at least twice. Therefore it can be compressed to  $a^4b^4$  so from 8 bytes down to 4 bytes if you encode it properly. On the other hand, if the input is highly mixed characters with few or no repetitions at all like abababab, the run length encoding of the string is  $a^1b^1a^1b^1a^1b^1a^1b^1$  which needs at least 16 bytes.

So the inherent problem with run length encoding is obviously the possible explosion in size, due to missing repetitions in the input string. Expanding the string to twice the original size is not really a good compression so one has to make sure the input data is fitted for rle as a compression scheme. One goal is to minimize the increase in size in the worst case scenario.

Also it should improve the compression ratio on data suited for run length encoding and perform better than the originally proposed rle.

## 1.3 Main Objective

Was ist das Ziel der Arbeit. Wie soll das Problem gelöst werden?

- bessere kompressionsrate im vergleich zu konventionellem RLE
- gleiche oder ähnliche decoding zeit ?
- ansatz beschreiben

The main objectives that derives from the problem statement is to archive an improved compression ratio compared to regular run length encoding. To unify the measurements, the compression ratio is calculated by encoding all files listed in the Calgary Corpus and then normalize the results.

Since most improvements like permutations on the input, for example a revertable Burros-Wheeler transformation to increase the number of consecutive Symbols or a different way of reading the bytestream take quite some time, encoding speed will increase. A secod objective might be to keep decoding speed close to the original run legth encoding.

## 1.4 Gliederung/Aufbau der Arbeit

Was enthalten die weiteren Kapitel? Wie ist die Arbeit aufgebaut? Welche Methodik wird verfolgt?

## 2. Basic principles of compression

Die Grundlagen müssen soweit beschrieben werden, dass ein Leser das Problem und die Problemlösung versteht, ohne weitere Literatur hinzuzuziehen.

basic knowledge:

- compression itself
- probability coding
- dictionary coding
- arithmetic coding
- lossy compression

To understand compression one first has to understand some basic principles of information theory like Entropy and different approaches to compress different types of data with different encoding and entropy. I will also show the key differences between probability coding and dictionary coding and a few comments on lossy compression.

### 2.1 Compression and Encoding fundamentals

...

### 2.2 Run Length Encoding

...

### 2.3 State of the Art

Die Literaturrecherche soll so vollständig wie möglich sein und bereits existierende relevante Ansätze (Verwandte Arbeiten / State of the Art / Stand der Technik) beschreiben bzw. kurz vorstellen. Es soll aufgezeigt werden, wo diese Ansätze Defizite aufweisen oder nicht anwendbar sind, z.B. weil sie von anderen Umgebungen oder Voraussetzungen ausgehen.

Je nach Art der Abschlussarbeit kann es auch sinnvoll sein, diesen Abschnitt in die Einleitung zu integrieren oder als eigenes Kapitel aufzuführen.

Beispiel, wie mit LaTeX zitiert werden kann: [? ? ?]



## 3. Analyse

In diesem Kapitel sollen zunächst das zu lösende Problem sowie die Anforderungen und die Randbedingungen einer Lösung beschrieben werden (eine präzisierte Aufgabenstellung).

### 3.1 Anforderungen

Anforderungen und Randbedingungen ...

### 3.2 Weiterer Abschnitt

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.



Abbildung 3.1: Logo der Universität Trier.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

Tabelle 3.1: Tabelle mit Werten.

A	B	C
Test 1	Slow	279
	Fast	499
	Very fast	719

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue dui dolore te feugait nulla facilisi.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis.

### 3.3 Zusammenfassung

Am Ende sollten ggf. die wichtigsten Ergebnisse nochmal in *einem* kurzen Absatz zusammengefasst werden.

## 4. Entwurf / Konzeption

In diesem Kapitel erfolgt die ausführliche Beschreibung des eigenen Lösungsansatzes. Dabei sollten Lösungsalternativen diskutiert und Entwurfsentscheidungen dargelegt werden.

To archive the main objective a few ideas to improve run length encoding were found. The first real difference is the conception of reading the data in chunks and then encode it vertically, meaning all the most significant bits in one chunk, then all second most significant bits and so on. The second change was to switch between the encoding of runs of characters bytes to count runs of ones and zeros so basically the same mechanism but on bit level.

### 4.1 Abschnitt 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

### 4.2 Abschnitt 2

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto

odio dignissim qui blandit praesent luptatum zzril delenit augue dui dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

## 4.3 Zusammenfassung

Am Ende sollten ggf. die wichtigsten Ergebnisse nochmal in *einem* kurzen Absatz zusammengefasst werden.

## 5. Implementierung

...

### 5.1 Abschnitt 1

...

### 5.2 Abschnitt 2

...

# 6. Evaluation

Hier erfolgt der Nachweis, dass das in Kapitel 4 entworfene Konzept funktioniert. Leistungsmessungen einer Implementierung werden immer gerne gesehen.

## 6.1 Abschnitt 1

...

## 6.2 Abschnitt 2

...

## 6.3 Zusammenfassung

Am Ende sollten ggf. die wichtigsten Ergebnisse nochmal in *einem* kurzen Absatz zusammengefasst werden.

## 7. Diskussion und Ausblick

(Keine Untergliederung mehr)

# Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich diese Bachelor-/Masterarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die aus fremden Quellen direkt oder indirekt übernommenen Gedanken als solche kenntlich gemacht habe. Die Arbeit habe ich bisher keinem anderen Prüfungsamt in gleicher oder vergleichbarer Form vor-gelegt. Sie wurde bisher auch nicht veröffentlicht.

Trier, den xx. Monat 20xx