Fortschritt-Berichte VDI

VDI

Reihe 10

Informatik/
Kommunikation

Nr. 864

Herwig Unger (Ed.)

Autonomous Systems 2019: An Almanac

FernUniversität in Hagen
Schriften zur Informations- und Kommunikationstechnik

Fortschritt-Berichte VDI

Reihe 10

Informatik/
Kommunikation

Herwig Unger (Ed.)

Nr. 864

Autonomous Systems 2019: An Almanac

FernUniversität in Hagen
Schriften zur Informations- und Kommunikationstechnik

Unger, Herwig (Ed.)
Autonomous Systems 2019: An Almanac
Fortschr.-Ber. VDI Reihe 10 Nr. 864. Düsseldorf: VDI Verlag 2019.
218 Seiten, 75 Bilder, 26 Tabellen.
ISBN 978-3-18-386410-2, ISSN 0178-9627,
€ 76,00/VDI-Mitgliederpreis € 68,40.

Keywords: Autonomous Systems – Natural Language Processing – Image Processing – Complex Systems – Machine Learning – Deep Learning – Simulation – Clustering – Safety – Ethical Aspects

Since 2010 it became a good tradition that colleagues from different research areas publish articles about their unsolved scientific problems, on-going work or research results in a joint book with the title "Autonomous Systems". Those systems exist in several areas of science, describing self-contained and self-controlled groups of possibly interacting or interrelated entities that form unified ensembles acting in given environment without outer control of any higher instances, leaders or managers.

In this year, three major topics are presented, which are valiant, strongly interfering with politics and everyone's discussion and are, therefore, rarely considered without emotions or political and financial concerns: the climate on our planet earth, ethics in the context of applying and using autonomous systems as well as problems of (mostly individual) road traffic. The following contributions deal with various topics related to theory, fundamentals, natural language and image processing as well as simulation.

Bibliographische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter www.dnb.de abrufbar.

Bibliographic information published by the Deutsche Bibliothek
(German National Library)
The Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at www.dnb.de.

Schriften zur Informations- und Kommunikationstechnik
Herausgeber:
Wolfgang A. Halang, ehemaliger Lehrstuhl für Informationstechnik
Herwig Unger, Lehrstuhl für Kommunikationsnetze
FernUniversität in Hagen

Als Manuskript gedruckt. Printed in Germany.
ISSN 0178-9627
ISBN 978-3-18-386410-2

Preface

Since 2010 it became a good tradition that colleagues from different research areas submit and publish articles about their unsolved scientific problems, ongoing work or research results in a joint book "Autonomous Systems" published by VDI-Verlag in Düsseldorf. When the editors were looking for a suitable title as a framework for these contributions they found that the term "autonomous systems" exists in several areas of science, describing self-contained and self-controlled groups of possibly interacting or interrelated entities that form unified ensembles acting in given environment without outer control of any higher instances, leaders or managers. The books prepared in the beginning as proceedings of the workshop and conference with the same title held each year on Majorca Island in October slowly changed their character to real, interdisciplinary almanacs. Now, authors can contribute to it without an obligation to participate or present their work at the conference. Moreover, the book is free of any reviews or censorship what shall motivate especially people thinking off the mainstream to submit their work. The book is distributed among the conference participants and shall be a motivation for free, unbiased discussion at the conference in Cala Millor. It shall remind everyone that dispute and competition of opinions is one, maybe the most important root of science, where decisions cannot be made by majority voting or compromises. Therefore, the search for the truth or the best, most efficient solution can even significantly be guided by the opinion or ideas of a single outsider or specialist in a completely other area of science, which may win over the majority of renowned specialists.

The editors are glad that the three keynotes presentation contained in this volume are touching three research fields, which are valiant, strongly interfering with politics and everyone's discussion and are, therefore, rarely considered without emotions or political and financial concerns: the climate on our planet earth, ethics in the context of applying and using autonomous systems as well as problems of (mostly individual) road traffic.

The following three chapters deal with natural language processing, fundamentals and theory as well as image processing. Although at first glance these topics seems to be very different, they are connected by two facts the authors emphasize. First, interfaces become increasingly human-like: people get used to program or to interact with machines in their natural languages, either in written

or in spoken form, and cameras are employed as eyes of or visual interfaces to computers. Secondly, as a consequence, the amount of data available in big data bases, the size of the systems to be managed as well as the computing performance needed are increasing rapidly.

As in the last years, the editor wants to say thank you to Mrs. Barbara Kleine and Mrs. Jutta Düring for their caring support in all technical and organisational details needed to finish this book and bring all authors and interested people together at our conference in Cala Millor. We hope that all of you will enjoy the results of our work and support us with your ideas and work in 2020 again.

Hagen, August 2019
Herwig Unger

Contents

Fundamentals and Theory

Image Processing

Keynotes

Is Global Warming Men-made? The Atmosphere Seen as an Autonomous System

Gerd K. Heinz[1]

Abstract: Today, it is a question of belief to find carbon dioxide (CO2) as the climate killer number one. Observing the field, we try interdisciplinary calculations and a qualitative *and* quantitative attempt to better understand the processes between natural sciences and economy. We find water vapour as the dominating greenhouse gas number one. The variation of solar radiation has absolute highest importance for global warming. Compared to water vapour, the rule played by CO2 is nearly negligible. CO2 follows the warming of the atmosphere as an indicator. We find desertification, following the population explosion of mankind as the second large heat- and CO2-source. This part of global warming is men-made. We find an sharp correspondence between growing air traffic and growing CO2 values. By contrast to the common sense we see that every CO2 production produces water vapour too, cooling or heating the earth atmosphere. Reducing the CO2-production will reduce water vapour production. This could increase the temperature of the atmosphere. So, the abandonment of nuclear- and coal power plants, together with the abandonment of oil and gas combustion by "green" technologies could heat the earth. By the way we calculate the economy and the risks.

1 For Future?

As a question of economic and financial survival of human civilization, global warming has got a story of the most explosive political significance. Lots of journalists, politicians, teachers and pupils know exactly the reason of climate change (if there is any): *Man-made carbon dioxide*. The Greta-Youth demonstrates

[1]The retired author studied electronics technology, mechanics and microelectronics at TU Dresden. He received the PhD in microelectronics at the HU Berlin. He made the first integrated circuit in the Berlin territory. Behind other inventions, he found interference networks able to calculate nerve nets. He applied this theory to invent the acoustic photo- and cinematography (Acoustic Cameras). He has no financial interests in any energy industries nor is he affiliated with any political party.

to stop the coal, natural gas and oil production immediately (and nuclear energy too). What they do not know: They demonstrate to stop our life. False assertions light the conflict.

But different scientist found other mechanisms for global warming to be much more important, then carbon dioxide.

700 years old barley corns (Gerste) in Greenland show, that the climate here was much warmer then today [1]. Under Alps glacier "Pasterze" a 600 years old tree came out [29], it was a larch trunk (Lärchenstamm). But this warm period was mutually not reasoned by men.

From the eruption of the Laki- and Grimsvötn-volcanoes 1783 – 1785 in Iceland we know, it brought megatons of CO2 into the atmosphere, but instead of warming up, the *earth cooled down*. Snow in the summer reduced the agriculture production in Western Europe. People were hungry, thousands starved. The French Revolution 1789 and the wars of Napoleon were the consequences.

Dependent of the interpretation of the data sources used (geological, whether records, tree rings, volcano eruptions, sunspots, solar radiance etc.), different researchers come to very different approximations of data concerning the influence of solar radiation and CO2 on temperature and climate, for example, see the 2015 overview of W. Soon et al. [34], see also [35, 36]. What they mostly forget, is the volcano story and the possibility of cooling.

In Germany, the CO2-themes got a *religious dimension*, as a note from the nationwide "Protestant Church Congress and Meeting" in Dortmund, June 19–23, 2019 demostrated. The organizer, Kirchentagspräsident Leyendecker, said: "*Wer nicht anerkennen will, daß der Klimawandel menschengemacht ist, hat beim Kirchentag nichts zu suchen*." (Who don't accept the men-made climatic change is not requested at the church's day).

Thinking about the Greta-Youth "Fridays for Future", we have to understand, that the demonstrations are *well organized by NGOs with political and financial interests*. They transports the pupils with busses, they install tribunes with microphones and loudspeakers. They print banners, flyers, bills and so on.

2 The Rule Played by the IPCC

In his 2006 film "*An Inconvenient Truth*" the former US-Presidential Candidate Al Gore had shown, that the climate change occurred over 600 000 years without influence by men. Based on a graphic chart he demonstrated the *closed correlation*

of temperature and CO2 over the last 600 000 *years*. His work in climate change activism earned him (jointly with the IPCC) the "Nobel Peace Prize 2007" [24].

What he "forgot" to name was (behind 8 other details, heard by a 2007 court case [24]), that the CO2-curve *follows* the temperature curve all the time typically *in a distance of 800 years*, as the paleo-climatologist Ian Clark remarked [14, 17].

So *CO2 is not the reason, it is the effect of climate change*. It is to expect, that the time constants of oceans causes that delay. Got Al Gore the Nobel-Price 2007 for a lie? However, he has lots of political and financial interests in the field, see [24].

Looking into different papers of the 1988 founded United Nations organization "*Intergovernmental Panel on Climate Change*" (IPCC) we find the medieval warm period completely removed as an "event of local evidence" [2]. Is IPCC an organization with other then scientific interests?

Al Gore tried to push the US-government 1997 into the Kyoto-Protocol [25]. He was opposed by the Senate, which passed unanimously (95/0) the "Byrd-Hagel Resolution", which stated, Kyoto "would result in *serious harm to the economy* of the United States" [24].

In other words: By unknown reasons in the history of earth the climate changed, but CO2 played no rule. So, CO2 seems not to be a potent "climate killer" gas? Other effects or mechanisms have very much more potential?

We note:

- The medieval warm-period (950 . . . 1300) was removed by IPCC
- The CO2 of volcanoes does not heat the atmosphere
- Temperatures varied all the time in earth history before men
- CO2 concentration *follows* temperature with 800 years delay

3 World Population and World Economy

Between 1960 (3 billion) and 2000 (6 billion) we had the *shortest doubling of population, mankind ever had*. Time for doubling the population becomes shorter and shorter. Each of us needs oxygen, heating material, food, we try to have clean water, electricity, a car, a flat or a house with roof, furniture, radio, television, internet; we produce gases and excrements. We need infrastructure, physicians, a supermarket, a bakery, a butcher, medicals, trains, busses, airplanes etc. .

It can be supposed, that the amount of energy we need and the amount of CO2 we produce, has a very closed relation (is proportional) to the exponentially growing population. Demands to remove the CO2 means in that consequence, to kill people? A growing number of wars in the world seems to indicate that.

Men need coal for steel and concrete production, oil for traffic, transportation and agriculture and gas for all kinds of heating. It is not possible, suddenly to switch over to other, unknown technologies. Neither we have the engineering knowledge to replace fossil energy per administrative command, nor we have the market economy for that attempt. And the costs by each technological change will explode.

Abrupt abandonment of fossil energy – as demanded by green parties around the world, or the "Fridays for Future" movement around Greta Thunberg, would replace industry, traffic and agriculture of mankind immediately by unknown technologies and even more, *by an unknown dictatorial planning system, excluding the market economy*. In Germany, we are on that way.

This could destroy civilization accepting the dead of million people worldwide. It could bring the biggest holocaust, mankind ever have seen. Electors, politicians or VIPs seem not to be clear, how dangerous it is, to play with this greenish fire. If we have removed industrial- and power plants, we do not have them anymore. If we remove the market economy, we lose each kind of democracy.

Last not least the size of the earth and the agricultural areas are constants and do not grow. Reasoned by population explosion, agricultural desertification grows up to a dangerous level. Desertification of giant areas each year becomes a growing problem witch influences the global warming process and the carbon dioxide production substantially, we will find.

4 Men-made Carbon Dioxide and Water Vapour

Burning oil or gas produces carbon dioxide and in the same volume water. Coal power stations and nuclear power stations produce giant amounts of water vapour by the cooling towers. Volcanoes showed us, that water has the priority over CO2: Water cools the climate down. In opposite to the general belief, our *fossil power generating technologies does not heat the atmosphere, they cool the atmosphere*.

To find out, how much water and carbon dioxide is man-made, Tab. 1 shows a calculation between chemistry, atom physics and energy production.

Table 1: World energy supply 2016, sources [3, 9]. One kilogram oil unit has by definition the energy of 11.63 kWh.

World energy 2016	Giga tons oil unit	%	C in %	CO2 in kg/kg	Giga tons CO2	H in %	H2O in kg/kg	Giga tons H2O
oil	4.418	33.3%	72.4%	2.654	11.73	27.6%	2.484	10.97
coal	3.732	28.1%	95.0%	3.483	13.00	5.0%	0.45	1.68
gas	3.204	24.1%	79.7%	2.922	9.36	20.3%	1.827	5.85
nuclear	0.592	4,5%						
hydro	0.910	6.9%						
renewable	0.419	3.2%						
sum	**13.275**	**100.0%**			**34.09**			**18.51**
Total energy	154.388	TWh						
Partial sum fossil	11.354	Giga tons ou			C ⇒ CO2	44/12	3.666	kg/kg
Total fossil energy	132.047	TWh			H ⇒ H2O	18/2	9	kg/kg

The columns "C in %" and "H in %" are raw approximations basing on the atomic weights ($H = 1$, $C = 12$, $O = 16$) and the not exactly known molecular structure, so the following columns are approximations too.

Reading example: Mankind produced in 2016 around 4.418 gigatons oil. Burning the oil, we get 11.73 gigatons CO2 and 10.97 gigatons water. Mankind burned in 2016 11.354 gigatons fossil energy, relating to an energy of 132.047 TWh (terawatt hours).

5 Scales and Units

To understand potential influences on climate change, it is not sufficient to discuss them verbally. We need scales to compare the potential effects of different heat sources. As a good scale the author uses since 2011 the world energy production (EW) [3], also known as "total power energy supply" (TPES).

Table 1 shows an EW of 154.388 TWh, this should be the heat, mankind produces at all per year. Important for the CO2 problem is also the fossil energy EF and the mass of CO2-production MCO2.

To calculate the corresponding, average power production we have to divide the energy by the number of hours of the year. 365 days $*$ 24 hours $=$ 8760 h/y. The energy of 154.4 TWh corresponds to an average power production of 154.4 TWh / 8760 h/y $=$ 0.0175 TW $=$ 17.5 GW. By analogy we get the total fossil power production with 132 TWh / 8760 h $=$ 15.1 GW.

"The scales"– important man-made values for 2016:

Energy production world	EW $=$ 154.4 TWh
Energy production fossil	EF $=$ 132 TWh
Power production world	PW $=$ 17.5 GW
Power production fossil	PF $=$ 15.1 GW
Man-made carbon mass	MC $=$ 34.1 / 3.66 $=$ 9.32 Gigatons
Man-made CO2 mass	MCO2 $=$ 34.1 Gigatons
Man-made H2O mass	MH2O $>>$ 18.5 Gigatons

On our scale, this values should have a (relative) scale of one.

6 The Greenhouse Effect

Wood and Nahle [10] have shown, that the main important effect of a greenhouse is not the radiation, it is the blockage of convection. They are totally right, but this is not the main question.

The important idea behind the greenhouse comparison is the spectral absorption of CO2 and H2O in the Mid-IR-range, Fig. 1. The electromagnetic wavelength corresponds inverse to the temperature. The range around 10 µm marks a temperature of 255 Kelvin $\sim -18\,^{\circ}$C, while the wavelength around 0.5 µm marks the incoming radiation of the sun. But the radiation energy is proportional to the frequency, that is inverse to the wavelength. The higher the frequency, the higher the energy. Incoming solar radiation has thousand times more energy, as the 10 µm lines, blocked by carbon dioxide. So the reflected energy is smaller, as higher the wavelength is, see Fig. 2. And the water absorption in the whole infrared range is much higher the CO2 absorption, to see in Fig. 1 and Fig. 2.

If we multiply the efficiencies of water vapour and CO2, we find the mid-IR back-radiation of water approximately 5-times stronger then CO2 [3]. And the atmospheric concentration of water can be up to 100-times higher as CO2 [3].

So *water is the very better greenhouse gas* with an up to $100 * 5 =$ *500-times higher efficiency then CO2*. And it is the greenhouse gas with the absolute highest amount

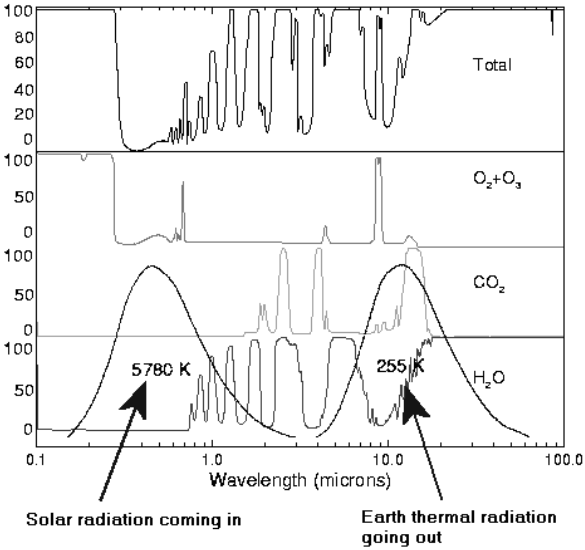


Fig. 1: Absorption spectra of water vapour, carbon dioxide and oxygen. We find oxygen blocking the UV-range, while water vapour and carbon dioxide blocks the cold back-radiation of the earth in the far infrared range. Image source [11].

in the atmosphere (by natural production). Looking into NASA-satellite observations [31], we see oceans and rain-forests as the big sources of water vapour.

We note, that *natural water vapour is in quantity and quality the most dangerous greenhouse gas*. It should urgently be remarked in an Annex to the Kyoto-protocol [25].

What the figure does not tell us: *water blocks the entire spectrum*, if it reaches the saturation point (also called dew point or condensation point). Although not known, this is the *most important point for all climate evaluations*. It is more important, as the whole rest of the figure. Why? If water blocks the incoming radiation completely, and the outgoing radiation too, any other "climate killer gas", like CO2, has no chance to modify the radiation anyway.

Hug [37] calculated 1998 the extinctions coefficient (absorption) of CO2 in a concentration of 357 ppm for a wavelength of 15 µm to 20.2 m^2/mol. Over a height difference of 10 meters follows an absorption of 99.94 % (practically all back-

radiation is stopped on 10 meters). The assumption is, that the atmosphere "is saturated" with CO2, any further CO2 will change nothing of interest.

We note:

- The troposphere does not work at all like a greenhouse
- Water has up to 500-times higher efficiency then CO2 as a greenhouse gas, it is by far the most productive greenhouse gas
- If it reaches the saturation point (clouds), water blocks the entire spectrum, giving other greenhouse gases no chance for any influence

7 Men-produced Gases Versus Gases in the Atmosphere

Seen as a ball, the earth has an average diameter of 12 730 km. The corresponding surface has 509e12 m^2 or 5.1e18 cm^2. With a pressure of approx. 1 kg/cm^2 the atmosphere has a weight of 5.1e18 kg = 5.1 Zg (zettagram) [3]. Wikipedia [5] says 5.15 Zg (*remark*: 1 Zg = 1e18 g = 1e12 tons = 1 Tera tons = 1000 Giga tons).

Natural water vapour in the atmosphere has a volume of 12 900 km^3 [32], the corresponding mass is 12.9 Teratons. With the mass of men-produced MH2O = 18.5 Gigatons we have a factor 12.9 Tt / 18.5 Gt = 697 $\sim$ 700. So men's influence to the natural water vapour cycle is $1/700 = 1.4$ promille.

Using the current CO2-concentration of 410 ppm (parts per million) we find *MPPM = 410 ppm * 5.15 Zg = 2.11 Gigatons (billion tons) CO2 in the atmosphere*.

To understand, if and how men influences the atmosphere, we have to ask for the relation between men-produced CO2 and CO2 in the atmosphere. The result surprises: *Men produces 16.2-times more CO2 per year as being currently stored in the atmosphere?* (MCO2 = 34.1 Gigatons; MPPM = 2.11 Gigatons). (It was one of the most confusing results, I ever had).

So we fill the atmosphere every 22 days (365 days/16.2 = 22 days) with CO2? What could this mean? The *atmosphere can not store our CO2*. It has to have very big, natural other resources for reduction. If not, the CO2-level of the atmosphere would be significant higher! It is an autonomous system.

The question is: Why isn't the CO2-concentration much higher? The answer can only be, that *oceans and forests bind the CO2 very fast*.

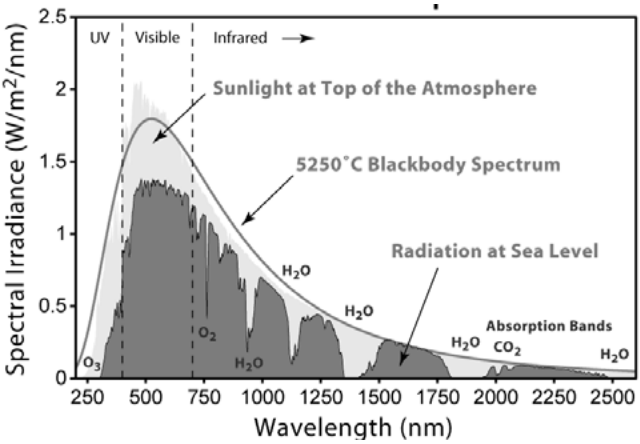


Fig. 2: Solar radiation spectrum. The absorption bands of CO2 are very small compared to H2O. If there is a greenhouse effect, it depends from water many times stronger then from CO2. Image source [8].

Different sources talk about a reduction of CO2 by natural processes in the range of 10 Teratons (10e12 tons) per year. Compared to 34.1 Gigatons men-made CO2 we find a factor of 10 Tt / 34.1 Gt = 294. So the *ability of nature to remove CO2 is estimated to be 300-times higher* the production of mankind [3, 6]. And CO2 has a three times higher density then air, it tries to sink to the bottom.

So CO2-sources near the ground (industry, traffic, power plants) have mutually no influence on the CO2-concentration in the troposphere.

It seems to be like a great wood-fire in our garden in deep winter. In the near it is hot, but the neighbor can not feel anything of the heat. The smoke reaches a height of 10 meter and we can smell the smoke only some hundred meters in wind direction.

We note:

- Because of the high molecular mass, CO2 sinks faster to the ground, while water vapour stays in the air for a long time
- Men produces per year 16-times more CO2, then atmosphere carries
- Nature produces 300-times more CO2, then men
- The atmosphere can not store our CO2

8 The Dual Rule Played by Water Vapour

Water vapour has a molecular mass ($16 + 2 * 1 = 18$) that is *comparable to air density* (oxygen 16, nitrogen 14). So, by contrast to carbon dioxide or other "climate killer gases", *air accumulates water vapour long times*. Only condensation stops the accumulation in form of rain or snow.

All other "climate killer gases" have higher densities then air, so they all sink to ground very fast, as the quantitative analyses of the man-made CO2 or of the Keeling-curve will show. By example, the density of CO2 ($12 + 2 * 16$) is three times higher, so CO2 sinks to ground comparable fast.

Fig. 1 and Fig. 2 show, that water acts for the back-radiation as a many times stronger greenhouse gas then CO2. Also it has a 100-times higher concentration up to 4 % and an nearly long surveillance in the atmosphere.

With the weight of the atmosphere of 5.15 Zg (zettagram) the total weight of water, that could theoretically be stored in the atmosphere is $4\,\% * 5.15$ Zg $=$ 0.206 Zg or 0.206e18 g or 206e15 g or 206 billion tons or 206 Eg (exagram).

This is potentially 100-times more as the current CO2 mass. Multiplied by a five times higher efficiency to warm up the atmosphere in the mid-IR back-radiation zone, the *total impact of water vapour can be approximately up to 500-times higher then CO2*.

So *water vapour is the most dangerous greenhouse gas on the earth?*

Water has a *second, much more important function* in the atmosphere. Sinks the temperature of air with a relative humidity of 100 % to the dew point, the water condenses, building clouds, that stops the solar radiation and the back-radiation radically, nearly complete. The earth cools down.

All big volcano eruptions brought strong falling temperatures reasoned by dark, cloudy sky and the blocking of sun radiation. They show, that the cooling down effect of water vapour in form of dark clouds is much more important, then the warming up by carbon dioxide or other "climate killer gases". *If clouds block the incoming solar radiation completely, the earth can not warm up!*

Burning oil or gas produces carbon dioxide and water vapour in comparable quantity [3]. Because of the density, water vapour has a higher time of survival in the atmosphere, so the efficiency to influence climate changes is much higher compared to CO2.

So, *global warming can be effected by the missing cooling by water vapour*, appearing as a *side product of all burning process of oil or gas*, or the *cooling process of coal power plants or nuclear power plants cooling towers*.

If we stop to use oil or gas, or if we stop the water vapour production of the cooling towers of energy plants, we reduce the accumulation of clouds with the effect of a higher incoming solar radiation. Growing temperatures could follow.

That means very clear: The *reduction of CO2 could bring us a higher global warming* because of the reduced water vapour and cloud generation.

We note:

- Water vapour has a 500-times stronger greenhouse effect then CO2
- Water vapour is the most dangerous greenhouse gas, we have
- Water vapour can block the incoming solar radiation
- If solar radiation is blocked by clouds, earth cools down
- All fossil burning- processes produce water vapour
- CO2 reduction will bring us a serious global warming problem

9 Men-produced Desertification

We know, that deserts have most of the time cloudless sky. In [3] the author calculated the influence of cloudy sky relative to cloudless sky.

Calculation for the area of the Sahara (source [3], actualized): The Sahara has a surface of 9 million km^2 = 9e12 m^2 (Wikipedia). We suggest rough a difference of 800 $Watt/m^2$ between cloudy and cloudless sky, 8 hours per day sunshine at 365 days per year, the energy difference ES is approximately ES = 9e12 m^2 $*$ 800 W/m^2 $*$ 365 d $*$ 8 h = 21e18 Wh/y = 21 EWh/y (Exa Watt hours per year).

Compared to the world energy production per year (EW = 154.4 TWh), the Sahara produces 21 EWh / 154.4 TWh = 136-times more heat, as men.

So for the area of the Sahara we find 2016 a 136-times higher warming potential, related to the world energy production (EW). All warm deserts together bring a warming energy of approximately 300 to 500-times the world energy production EW. This is very much more, then all man-made effects can produce together!

Thinking about men-produced desertification [3], we find other important sources for warming-up.

Between 1960 (3 billion) and 2000 (6 billion) we had the shortest doubling of population, mankind ever had, Fig. 3, [26]. Time for doubling becomes shorter and shorter. Mathematicians know, this process is called exponential growing. Exponential functions describe any kind of explosion processes. So we can call it "population explosion" [26].

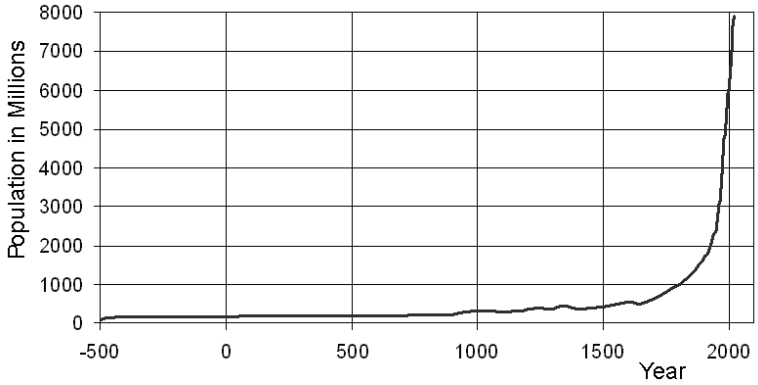


Fig. 3: World population explosion over the time [26].

Leading cities are Casa Blanca and Teheran, the population exploded here in 100 years by a factor of 100. The patterns are comparable: A village needs firewood, range land and acres. The land is cleared from forests. Depletion and desertification follows. Simultaneously the village grows to a concreted city inside a desert. We find these pattern in thousands of cities, from Syria to Afghanistan, from Morocco to Yemen, from Argentina to Mexico.

At the same level, the population increases, the *world energy consumption, the heat production, the water vapour and CO2-production have to increase*.

But the surface of the earth has a constant size. Areas for agriculture do not grow with the explosive growing society. More and more people need something to eat. So the efficiency of agriculture must grow, for example by the use of insecticides, fertilization and industrial animal husbandry.

At the other hand, men try to occupy forbidden areas for agriculture production, for example in rain-forests [3]. The permanent removal of rain-forest, the expansion of mega-cities, the over-production on agricultural areas and the succeeding, hermetic sealing of the ground by traffic infrastructure, buildings and roofs (also by solar-panel fields) reduce the natural ability to produce water vapour. IPCC says [2], an area of the size of *Germans agriculture area (0.12 km^2) dries out each year*. Rain-forests change to deserts.

If 0.12 million km^2 rain forest becomes a desert each year, this is 0.12 mio km^2 / 9 mio km^2 $= 1/75$ of the Sahara area. Drying-out, two things appear with relevance to climate change:

First, the cloudless sky over a desert warms up the earth with about *2 EW more ($1/75 * 136 \sim 2$) each year* (EW: world energy production of mankind). Each year two EW more means, *every 75 years we have a further Sahara, heating the earth* [3].

Second, if the rain forest's carbon is equivalent to a 50 cm thick layer, it produces *up to 25-times the mass of the men-made CO2-production* (MCO2) per year [3]. So, both effects influence the climate substantially [3].

We note:

- Sahara warms up the earth 136-times more then men (EW)
- All warm deserts together produce 300- to 500-times the EW
- Earth population grows explosive, producing growing desertification
- Desertification dries out 0.12 million km each year
- Desertification produces 2-times EW more heat per year
- Desertification produces 25-times more CO2 as all fossil sources

10 The Influence of Solar Radiation: Black Spots on the Sun

Black spots are heavy explosions on the sun surface caused by magnetic fields. While the material blows out, it cools down. So we find this explosions as black spots. By changing black spots activity, the solar radiation can vary up to five percent [3, 27]. Which energy is behind such a variation, Fig. 4?

Seeing the earth as a plate perpendicular to the sun, it has a surface of 510 million km^2. The solar constant is 1366 W/m^2 (Watt per square meter). Suggesting, clouds on the earth reflect 50% back into orbit, the other 50% reach the surface.

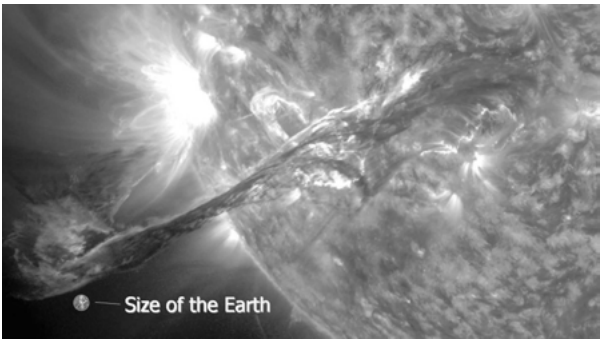


Fig. 4: A solar flare of the sun has many times the size of the earth, source [27]

The total incoming energy is supposed to 50 % $*$ 510 million km^2 $*$ 1366 W/m^2 = 348 PW (Petawatt). For a sun radiation variation of 5 % we get a power variation of 348 PW $*$ 5 % = 17.4 TW.

5 % solar power variation	PV = 17.4 TW
Power production World	PW = 17.5 GW

Dividing the solar radiation variation by the world power production, we get a factor of 17.4 TW / 17.5 GW = 995, Wikipedia says 10891-times more [33].

That means, a 5 % *variation of solar radiation has an effect, that is 1000-times higher as the world energy production* (EW). Thus the variation of solar radiation is found to be the most important effect on climate changes.

Fig. 5 shows measuring results of Shaviv [21], substantiating the calculation, that the sun has highest impact on temperature variation on earth. Suggestion for Fig. 5 was the ice smelting at the poles and the correlation with earth temperature. As more ice smelts, as more the sea level increases.

Observing black spots and the solar constant, different astro-physicists registered a stronger solar activity *within the last 70 years*. The activity is as high, as in the Holocene warm period 8000 years ago.

It is known, that high solar radiation brings the water vapour over the dew point. Clouds tend to disappear with higher solar radiation. We know the effect, if we observe the sky in hot regions in the early morning. Clouds disappear fast.

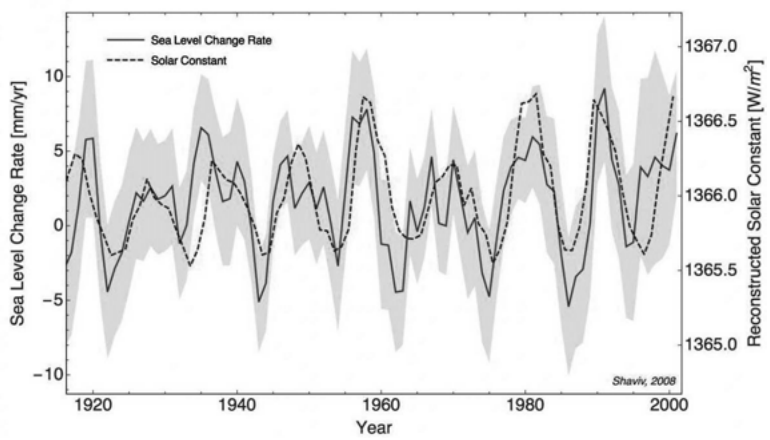


Fig. 5: The relation between solar radiation and sea level, source [23]. Red Dots: solar constant, blue line: sea level variation

Dependent on strength, flares have velocities between 1200 and 300 km/sec [22]. With a sun-earth distance of 149.6 million km they arrive earth between 34 and 138 hours. Because a solar flare reaches the earth after days, Dr. Piers Corbyn started to use that knowledge for the long-term weather forecast, see [14] at 27:45. And he has success.

What does it mean? It means not more and not less, as the *solar radiation causes the climate change* most substantially.

11 Emission by Airplanes

Gases with different densities try to split into fractions, where the weighty gas sinks down. CO2 has three times the density of air, water has that of air.

If the atmosphere would integrate over ground-near CO2, we should find coincidences between the atmospheric CO2-content (Keeling-curve) [7] an the fossil energy production (EF) on men by year. But we can not find coincidences, the *Keeling-curve does not represent the EF*.

Airplane emissions occur up to a high of 11 km (36 000 ft). At open sky, sometimes we see the condensation trails (contrails). We know the long, high and diffuse cloud-figures of the type *cirrus aeroplanus*.

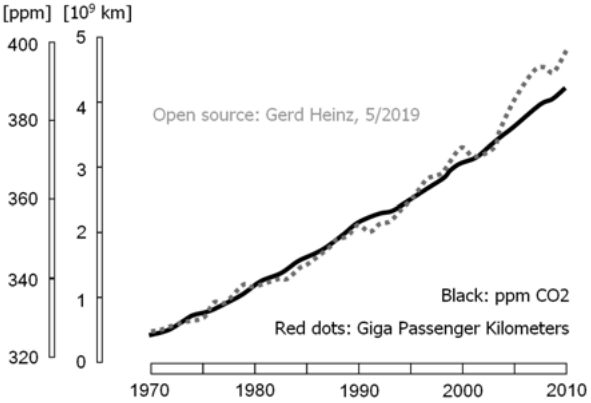


Fig. 6: Only coincidence? Keeling's CO2 measure on Mauna Loa (black) and total passenger kilometers of airplanes worldwide. The kerosene consumption of modern airplanes decreases, so they produce mutually after the year 2000 more passenger kilometers with a decreasing CO2 rate, source [3].

Keeling's research institute was on the Mauna Loa (Hawai) in a height of 3400 meter. Hawai has no important industrial infrastructure. Where should the CO2 come from? Maybe from airplanes? The idea appeared stupid, but I tested it in May 2019. I was surprised. If we compare the Keeling curve with the total billion passenger kilometers of airplanes over the years, we find a clear correspondence, see Fig. 6.

The total amount of CO2 and water produced by airplanes has a volume of 1.8 Gigatons, whereof CO2 has 1.08 Gigatons, water has 0.672 Gigatons [3]. In comparison to the amount of CO2 in the atmosphere (2.11 Gigatons) *this is the half!* What could it mean? The gas has a long way to come back to earth. This CO2 *can influence the back-radiation* and so substantially a climate change.

But in relation to desertification, producing 25-times the men-made fossil CO2-mass each year, airplane emissions are comparable very small, the amount of CO2 is $1/(25 * 2) = 1/50 \sim$ *20 promille of the volume of desertification*.

Natural water vapour in the atmosphere has a volume of 12 900 km^3 [32], the corresponding mass is 12.9 Teratons. We have seen, airplanes produce about 0.672 Gigatons. This is a factor 12.9 Tt / 0.672 Gt = 19 196 $\sim$ 20 000. So air traffic has only an influence of $1/19\,196 =$ *0.05 promille compared to natural water in the atmosphere*. This is not much. For further reading, see [28, 38].

It seems, the *sun heats the atmosphere since 70 years a bit more*, observable by higher black spot activity. By pure coincidence, we have *a growing CO2-concentration reasoned by growing air traffic* since 1970. And men-made desertification heats the earth very heavy.

So climate change is influenced by airplanes? With a small probability it is possible by the accumulation of CO2 and water vapour in the height of 10 km.

12 Germany's Energy Revolution called "Energiewende"

Oil, gas and coal reserves of the earth are limited. So the "Energiewende" appears as a good idea. But by a closer view, a fast introduction appears as an economic disaster, removing any market economy. The high volatility and the high prices of energy delivery by solar and wind power stations cause problems [7] as much, as necessary restrictions against customers.

Germanys contribution can be estimated: The men-made fossil mass of CO2 per year is 13 105 million tons oil unit. If Germany produces a volume of 317.8 million tons oil unit [9], Germany's contribution is 317.8 / 13 105 $= 0.024 \sim 24$ promille.

If the natural CO2-resources are 300-times bigger the men-made fossil mass of CO2 per year, *Germany's contribution sinks down to 24 promille / 300 = 0,081 promille*, related to all CO2 sources together. For that reason, we are about to destroy our industry and economy.

Although known by lots of scientists, that the atmosphere is not influenced by our CO2 substantially [35, 36] and that fossil carbon dioxide plays no substantial rule for the climate, Germany has started a revolution, that remembers to Kaiser Wilhelm's words "Am deutschen Wesen soll die Welt genesen!".

Germany decided after the Fukushima accident 2011 to abandon all 17 nuclear power plants until 2022. By end of 2015, nine were shut down [11]. 2011 they produced 20 % of the electric power of Germany [7].

Moreover, Germany announced for 2020, to reduce the CO2 emissions by 40 % versus 1990, especially by the reduction of coal-generated power [7, 18].

Coal power stations gets the new task, to compensate the deficits of energy, caused by volatility of wind power and solar power. Thus they have to produce on demand, increasing their costs per kWh to more then 200 % [7].

Because they are not able to deliver for current market prices, solar and wind power plants can only survive, if they deliver to their (much higher) production costs. Although hidden by subventions, this doubles the prices for energy production too.

Energy buffers are needed. The necessary, permanent storage and release of energy by pumped-storage stations consumes a substantial part of energy, to pay by the customer.

Taking all together, the total energy costs (for electricity) will increase rough by a factor of 500 %, politically hidden by subventions that are tax-paid.

To reduce the pump-station storage volume and so the costs, the customer is forced to buy new devices (washing machines, tumble dryers, dishwashers, heatings, E-cars), that have to be remote-controlled via internet by the energy-delivering companies. Regulated by energy costs, they tell the device, if energy is available. So the devices can wait some days, before they work.

Not only the energy production industry is named. Also the people. Government introduced hundreds of restrictions *to destroy the market economy* in perfection. By state-restricted economy and the *elimination of competition* the *customer is forced to buy high expensive things, that would not survive at a free market*.

For example: If I build a new house, government rules restrict the materials and devices I have to use. I have to take only energy sources with renewable energy (for example very expensive air- or groove-water heat pumps) that have prices and running costs far away from a simple gas-heating. I have to use special isolated types of windows, the isolation of the building has to meet regulations, the hourly prices for solar and wind energies are regulated and so on.

All this needs tax-generated subventions. The *German government switches off the principles of the successful market economy* of the 1950th. 30 years after the financial and economic ruin of the GDR-Planwirtschaft, the *German government substitutes the market economy by the next, state-restricted, GDR-like economy*. And all this, without of any reason. The GDR-propaganda called this: "Überholen, ohne einzuholen!" (passing by without reaching).

It needs no intelligence, that in consequence the last sensitive or power intensive production lines will leave Germany or they plan to escape. Without words: This is a very dangerous experiment. But the count down is running, and we all have to pay the price.

13 E-Mobiles Produce More CO2 than Combustion Mobiles

E-cars need to be charged. If they are charged at night, they can only use pump-stored solar energy or wind energy. Solar and wind energy cost more as the today's energy. If there is not enough solar- or wind-energy, E-cars will be charged by coal. If they are charged nightly, they use pump-stored solar- or wind-energy or coal energy.

Diesel- or Gasoline engine needs only one combustion step with a efficiency by 40 %, the coal power plant needs also this step with a comparable degree of efficiency. But 15 further steps follow to load the E-car's accumulator, if the energy is stored by a pump-station:

> *The primary electricity will by transformed to medium-level voltage. It will be transformed to high-level voltage. It will be carried by a long transmission line to the pump station. It will be transformed to the motor/generators voltage. It pumps the water up. The water falls down. The turbine with motor/generator produces voltage. It will by transformed to medium-level voltage. It will be transformed to high-level voltage. It will be carried by a long transmission line to the consumer side. It will by transformed to medium-level voltage. It will by transformed to power supply voltage (230/400 Volt). By the charge station it will be converted to cars voltage. It loads the accumulator. The accumulator gives its charge to the motor-controller of the car. Finally the E-motor produces the mechanical energy!*

Every of these fifteen steps has a limited efficiency, with maybe 5 % to 15 % energy loss per stage, that is transformed into heat. Multiplying the partial efficiencies, we get the overall degrees of efficiencies:

$$\begin{aligned} \text{Averaged efficiency } 95\,\%\text{: } (0.95)^{15} &= 0.46 \\ \text{Averaged efficiency } 90\,\%\text{: } (0.90)^{15} &= 0.21 \\ \text{Averaged efficiency } 85\,\%\text{: } (0.85)^{15} &= 0.087 \end{aligned}$$

What does it mean? In the best case, we make 46 % of the coal energy to mechanical energy, driving the E-car. In the worst case we get 8.7 % efficiency compared to a Diesel- or Gasoline engine. 54 % respective 91 % heat the atmosphere!

Using coal power to load E-cars means, we produce something between 2 to 10-times more CO2 in comparison to combustion engines to produce the same mechanical energy. This is terrible and not green. So it is not productive, to use coal power to

load E-mobiles anyway. In addition we get five times higher prices for electrical energy, compared to today's technologies.

In addition, the production of batteries and E-motors produce unknown masses of CO2, so different scientist [7, 35, 36] calculate, that E-cars run more effective as an Diesel, if they have a *lifetime over 300 000 to 600 000 kilometres*. What they forget:

1) The batteries have mutually a very shorter lifetime.

2) E-cars are mobiles for short distances. It is far to be realistic, that they will bring such high kilometre accounts in average.

So the *E-mobility concept will produce much more CO2, then current technologies*.

The EU-legislation "Reduction in CO2 emissions of new passenger cars" "gives super credits as incentives given to manufacturers to register low-emitting cars" [30]. These are not E-cars, if they are charged by fossil energy.

So the German "Energiewende" will become a technological, technical, economical and financial disaster only comparable with the destruction of industry and economy by World War II. It will remove the successful "Marktwirtschaft" (market economy) by a dictatorial GDR-type "Planwirtschaft" (planning economy). Last not least, the exploding prices can inspire heavy protests by poor people, bringing the next *political revolution*.

14 Summary

Volcanoes show, that the cooling effect of cloud-building water dominates over all other greenhouse gases.

The density of non-saturated water vapour is comparable to air, so it stays in the air for a long time, while all other greenhouse-gases (like CO2 or condensing water vapour) have much higher densities and sink faster down.

Water is the *most dangerous "climate killer gas"* at all, with a 500-times higher potential then CO2. It plays a *dual rule for cooling and heating* the atmosphere. In condensed form (clouds) it cools down, in weak- or non-condensed form it can isolate or heat the earth. Men's influence to the natural water vapour cycle is $1/700 = 1.4$ *promille*. It is negligible.

Dry regions do not reduce CO2, compared to wet regions of the earth. So CO2 *follows the warming of earth as an indicator*. If CO2 would also be the source, we

would get a feedback system, producing a *self-excitation*. Because this is not the case, and it is clear, that CO2 is an indicator, we can follow, that *the thesis: "CO2 is the source for global warming" can not be true*.

Nature produces 300-times more CO2, then men. But the yearly *CO2-production of men is 16-times higher the actual CO2-mass in the atmosphere*. So the *atmosphere removes all ground-near natural and men-produced CO2 very fast*. Both effects have no coincidence with Keeling's CO2-measures.

Compared to the world energy consumption EP, desertification produces each year *two times more heat, producing every 75 years the next Sahara! Compared to fossil, men-made CO2 it produces up to 25-times the mass. Desertification by population explosion has the highest men-made influence on climate change.*

So we need programs, to stop the population explosion, which provokes desertification.

With 1000-times of the world energy production (EW) by far the *highest influence on global warming has a 5 % variation of sun radiation*, registered by a higher amount of black spots on the sun surface since 70 years. This part of global warming is not men-made.

Combustion of fossil substances plays a dual rule. Combustion of oil or gas produces CO2 together with nearly the same amount of water vapour. But the earth atmosphere near ground is saturated with CO2 [37]. At a high of 10 meters we find a absorption of 99.94 % reasoned by CO2.

Airplanes produce only 0.05 promille of the natural water of the atmosphere. This is not important. But airplanes produce per year *the half of the measurable atmospheric CO2*. And *we find growing CO2-concentrations as the direct indicator for growing air traffic*. This can mean, that CO2 sinks comparable fast down to the earth. And it means too, this high CO2-concentration together with water vapour of airplanes in great height can create a *second isolating roof around the earth* and can influence global warming. If we would stop the air traffic in the height of 10 km today, we could have in two years a CO2-concentration with the level of 1970 on Mauna Loa.

Because all the time the CO2 appears in closed combination with water vapour, any combustion of fossil energy can inspire clouds, cooling the earth down, instead of warming it up.

Compared with Diesel-engines, E-cars produce up to 10-times more CO2, if they are charged with energy from gas- or coal-power plants.

Conclusion: We find solar radiation variation as the most dangerous (natural) source for climate change. We find desertification by population explosion as the most important source for men-made climate change. If we'd like to do something for the climate, we have to stop the population explosion now! Last not least, a second "greenhouse roof" made by airplane traffic can be possible.

References

[1] https://en.wikipedia.org/wiki/Greenland

[2] https://en.wikipedia.org/wiki/IPCC

[3] Heinz, G.: Klimawandel – Erderwärmung erzeugt Kohlendioxid. http://www.gheinz.de/publications/papers/2011_KlimaundKohlekraft.htm

[4] https://en.wikipedia.org/wiki/World_energy_consumption

[5] https://de.wikipedia.org/wiki/Erdatmosphaere

[6] https://en.wikipedia.org/wiki/Carbon_dioxide (includes the Keeling-curve)

[7] Sinn, H.W.: Buffering Volatility: A Study on the Limits of Germany's Energy Revolution. Energy and Climate Economics. Cesifo working paper no. 5950, version of May 2017, ISSN 2364-1428

[8] Solar radiation spectrum see https://en.wikipedia.org/wiki/Sunlight. image copyrigths by Nick84 - File:Solar_spectrum_ita.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24648395

[9] http://www.bp.com/content/dam/bp/en/corporate/pdf/energy-economics/statistical-review-2017/bp-statistical-review-of-world-energy-2017-full-report.pdf. pp. 9ff.

[10] Nahle, N.S.: Repeatability of Professor Robert W. Wood's 1909 experiment on the Theory of the Greenhouse. July 5, 2011. Biology Cabinet Online-Academic Resources and Principia Sci. International, Monterrey, http://www.biocab.org/Experiment_on_Greenhouses__Effect.pdf

[11] Sabhlok, S.: The science behind global warming and its impacts. A compilation of questions and answers. http://sanjeev.sabhlokcity.com/co2/

[12] https://de.wikipedia.org/wiki/Charles_David_Keeling, Atmospheric CO2 Data – Primary Mauna Loa CO2 Record, source: R. F. Keeling, S. J. Walker, S. C. Piper and A. F. Bollenbacher "Scripps CO2 Program". scrippsco2.ucsd.edu/assets/data/atmospheric/stations/in_situ_co2/monthly/monthly_in_situ_co2_mlo.csv

[13] ICAO: The world Aviation – 1950 to 2012. Revenue Passenger Kilometers: https://www.icao.int/sustainability/Pages/Facts-Figures_WorldEconomyData.aspx (20.5.2019)

[14] The Great Global Warming Swindle. British TV-documentation of March 8, 2007 on Channel4. Al Gore at minute 18:25, Ian Clark behind: https://www.youtube.com/watch?v=oYhCQv5tNsQ&feature=youtu.be

[15] RTL EXTRA spezial: DER KLIMASCHWINDEL – DOKU. Kurzfassung von [14], https://www.youtube.com/watch?v=RqWIN3Wc5tg (27.6.2019)

[16] Kipp, Rudolf: Beispiellose Erwärmung oder beispiellose Datenmanipulation? http://www.science-skeptical.de/blog/beispiellose-erwarmung-oder-beispiellose-datenmanipulation/001195/ (20.11.2009)

[17] Solanki S. K., Usoskin I. G., Kromer B., Schüssler M. , Beer J.: Unusual activity of the Sun during recent decades compared to the previous 11000 years. Nature, volume 431, pages 1084–1087 (2004) https://www.nature.com/articles/nature02995

[18] Bundesministerium für Wirtschaft (BMWi): Zahlen und Fakten Energiedaten, https://de.wikipedia.org/wiki/Energieverbrauch (23.09.2018)

[19] https://klimakatastrophe.wordpress.com/2019/04/27/das-geschaeftsmodell-fridays-for-future-der-plant-for-the-planet-foundation-des-club-of-rome/ (28.6.2019)

[20] https://klimakatastrophe.wordpress.com/2009/02/26/co2-absorption-im-selbstversuch/ (28.6.2019)

[21] Archibald, D.: Saltbush Solar Activity Watch. http://saltbushclub.com/wp-content/uploads/2018/12/solar-activity.pdf

[22] Climate Conversations. Dr. Ian Clark, Dept. Earth Sci., Univ. Ottawa: https://www.youtube.com/watch?v=UhNSQLvtRaE (10.3.2016)

[23] Nir Shaviv's blog: My experience at the German Bundestag's Environment Committee in a pre-COP24 discussion (2018). http://www.sciencebits.com/bundestag, Talk at Youtube: Wissenschaftler entlarvt Klimaschwindel in Bundestag. https://www.youtube.com/watch?v=xJSVjCrrCu0

[24] https://en.wikipedia.org/wiki/Al_Gore

[25] https://en.m.wikipedia.org/wiki/Kyoto_Protocol

[26] https://de.wikipedia.org/wiki/Bev\unhbox \voidb@x \bgroup \accent 127o\egroup lkerungsexplosion, data source https://www.census.gov/ipc/www/worldhis.html

[27] https://en.m.wikipedia.org/wiki/Solar_flare, image source NASA Goddard Space Flight Center, CC BY 2.0, adds by the author

[28] Bock, Lisa and Burkhardt, Ulrike: Contrail cirrus radiative forcing for future air traffic. Atmos. Chem. Phys., ACP 19, 8163-8174, https://doi.org/10.5194/acp-19-8163-2019, 2019

[29] https://klimakatastrophe.wordpress.com/2008/09/08/die-alpengletscher-und-der-eismann-otzi/ (30.6.2019)

[30] EU-legislation "Reduction in CO2 emissions of new passenger cars". https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM%3Ami0046

[31] https://earthobservatory.nasa.gov/global-maps/CERES_NETFLUX_M/MYDAL2_M_SKY_WV

[32] http://www.pci.tu-bs.de/aggericke/PC5-Atmos/Wasser.pdf

[33] https://de.wikipedia.org/wiki/Strahlungshaushalt_der_Erde

[34] Soon,.W., Connolly, R.+ M.: Re-evaluating the role of solar variability on Northern Hemisphere temperature trends since the 19th century. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.729.6404&rep=rep1&type=pdf

[35] http://www.schmanck.de/Klimadeutsch.html

[36] https://www.eike-klima-energie.eu/2015/05/17/22-sehr-unbequeme-klima-wahrheiten/

[37] Hug, Heinz: Die Klimakatastrophe – ein spektroskopischer Artefakt? http://uploader.wuerzburg.de/mm-physik/klima/artefact.htm

[38] Dlugokencky, E. et al.: Observing water vapour. WMO-bulletin 65(2) 2016. https://public.wmo.int/en/resources/bulletin/observing-water-vapour

Ethical Aspects of Autonomous Systems

Gudrun Schiedermeier

Faculty of Computer Science, University of Applied Sciences, Landshut, Germany

Abstract: Recent developments in Autonomous Systems such as self-driving cars, service-robots, Lethal Autonomous Weapon Systems (LAWS) or software agents (bots) including Artificial Intelligence (AI) raise complex moral questions. Who is responsible for possible harm? What are the moral, social and legal consequences of their actions? Should some of those systems even be banned? Who controls them? Who regulates the design, use and development? There seems to exist a broad consensus, that these systems should respect human and civil rights, that some frameworks to guide design and development of autonomous systems are necessary and that a meaningful human control is essential for moral responsibility. There are several initiatives from researchers, practitioners, various groups and professional associations, such as the IEEE initiative on Ethics of Autonomous Systems [1], the Statement on Artificial Intelligence, Robotics and Autonomous Systems from the European Group on Ethics in Science and New Technologies [2] or the Asilomar AI Principles [3]. Nevertheless, a dialogue is recommended that "focuses on the values around which we want to organise society and on the role that technologies should play in it." [2]

1 Introduction

Advances in robotics, autonomous vehicles or AI in Google's Deep Learning Software have generated a lot of interest from the broad public. People around the world use systems like Amazon's Alexa on a daily basis. Robots from Boston Dynamics such as Atlas move elegantly over obstacles and are able to perform backsaltos [20]. According to Californias Department of Motor Vehicles Waymo self-driving vehicles drove more than 1.2 million miles in California, most without human operators intervening [4]. Although Uber's fatal crash in Tempe, Arizona in 2019 killed a pedestrian, the company was not liable, Uber won't be charged with a crime [5]. In 1997 IBM's Deep Blue already beat Garry Kasparov

at chess and in 2016 Google's AlphaGo beat the human Go World Champion. AlphaZero, Google's latest AI system, didn't learn from human beings – for the first time ever – it bootstrapped itself to world champion level, having been fed with the rules of chess and playing against itself [6]. Often it is impossible to understand the system as a whole or the parameters of the underlying algorithm and it is not transparent how decisions were made. Some intelligent software systems like bots or robotic systems are released into the world unsupervised. Humans might not foresee their actions.

2 Definitions

"Conventional *automation systems* enable low-level processes to run without human intervention under normal conditions. Human decisions are still required for more complex tasks." ([7]) Making automation systems more autonomous means handing over an increasing number of these tasks to the system. Autonomous Systems require multiple sensor systems including cameras, lidar, radar and other sources producing huge amount of data. Quality of data is essential, because a first selection of data may already change the overall result. "Artificial intelligence is a valuable technology for processing this data. AI is not the same as autonomous systems. AI is a technological means to achieve a specific level of autonomy. Autonomy is the goal that AI can help achieve." ([7]) AI and especially machine learning is required to model the complexity of real-world environments.

DARPA (Defense Advanced Research Projects Agency) describes autonomous systems as "Systems that are aware and interact with their environment". It defines intelligent systems as "systems that know what they're doing" and "exhibit the following abilities:

They

- will be able to infer and reason, using substantial amounts of appropriately represented knowledge.
- will learn from their experiences and improve their performance over time.
- will be capable of explaining themselves and taking naturally expressed direction from humans.
- will be aware of themselves and able to reflect on their own behavior.

- will be able to respond robustly to surprises and explore in a very general way.

- will be able to interact/interface with humans . . . ." [8]

3 Types and Levels of Autonomous Systems

According to DARPA there are the following types of Autonomous Systems [8]:

- Unmanned Vehicles: Undersea, on sea, on land (driverless cars), in the air and in space. For these systems energy efficient locomotion, navigation and mobility in various environments is important.

- Robots: Perception, planning and decision making as well as mobile manipulation is relevant.

- Swarms: Collection of autonomous systems with distributed communication and control. Communication between individual agents and high-level human control, adaption, learning and reconfigurability are significant.

The European Group of Ethics encompasses [2] self-driving cars and drones, robots in deep sea and space exploration, weapon systems, software agents, such as bots in financial trade and deep learning in medical diagnosis as examples of autonomous technology and artificial intelligence.

Thomas Gamer suggests [. . .] six levels of autonomy [7] with an increasing independency from humans from 0 (without any autonomy) to 5 (complete absence of humans) (Table 1).

4 Key Questions and Considerations

According to the European Group on Ethics [2], systems that operate increasingly independently of humans, raise five questions:

1. Questions about safety, security and the prevention of harm: "How can we make a world with interconnected AI and autonomous devices safe and secure and how can we gauge the risks?"

Table 1: Levels of autonomy by Thomas Gamer

0	No autonomy, humans are in complete control without assistance.
1	Assistance with or control of subtasks. Humans always responsible, specifying set points.
2	Occasional autonomy in certain situations. Humans always responsible, specifying intent. Prerequisite Automation system monitors the environment.
3	Limited autonomy in certain situations. System alerts to issues. Humans confirm proposed solutions or act as fallback.
4	System in full control in certain situations. Humans might supervise.
5	Autonomous operation in all situations. Humans may be completely absent.

2. Questions about human moral and responsibility: "Where is the morally relevant agency located in dynamic and complex socio-technical systems with advanced AI and robotic components?" "Does it make sense to speak about 'shared control' and 'shared responsibility' between humans and smart machines?"

3. Questions about governance, regulation, design, development, inspection, monitoring, testing and certification: "How should our institutions and laws be redesigned to make them serve the welfare of individuals and society and to make society safe for this technology?"

4. "Questions regarding democratic decision making, including decision making about institutions, policies and values that underpin all of the questions above."

5. Questions about explainability and transparency of AI and autonomous systems: "Which values do these systems effectively and demonstrably serve?" "Around which values do we want to organise our societies?"

The term 'autonomy' goes back to philosophy, especially Kant. "Autonomy is an individuals capacity for self-determination or self-governance. Moral autonomy, usually traced back to Kant, is the capacity to deliberate and to give oneself the moral law, rather than merely heeding the injunctions of others." ([9])

In this original ethical sense "no [...] system can in and of itself be called 'autonomous' ", because "they cannot be accorded the moral standing of the human person and inherit human dignity" [2]. The terminology 'autonomous system'

here refers to "the highest degree of automation and highest degree of independence from human beings in terms of operational and decisional 'autonomy' " [2].

5 Ethical Framework

Designers, engineers and computer scientists shouldn't forget about human aspects in autonomous systems. Obviously, it doesn't make sense to set up a framework of constraints that everybody must follow. On the other hand it must be clear where and to what purpose systems will be deployed. For example, the artificial intelligence of an industrial robot operating in a production line must differ greatly from the AI of a robot in health-care, which requires more strictness in terms of environment sensitivity [10]. We need rules how systems should be programmed for "morally acceptable outcomes in terms of lives lost and, respectively, lives saved" [2].

- In 2017 the former German minister of transport, Alexander Dobrindt, installed a committee on ethical questions regarding driverless cars. The outcome was, in short, that autonomous vehicles can't solve a serious moral dilemma like the 'Trolley-Problem'. Neither the programmer nor the machine should be allowed to decide about life or death of human beings [13].
- An Open Letter from AI & Robotics Researchers was announced on July 28 at the opening of the International Joint Conference on Artificial Intelligence 2015, signed among others by Stephen Hawking, Elon Musk, Steve Wozniak and Juergen Schmidhuber. They strictly advised against using AI to build autonomous weapons. "Many arguments have been made for and against autonomous weapons, for example that replacing human soldiers by machines is good by reducing casualties for the owner but bad by thereby lowering the threshold for going to battle. The key question for humanity today is whether to start a global AI arms race or to prevent it from starting." "In summary, we believe that AI has great potential to benefit humanity in many ways, and that the goal of the field should be to do so. Starting a military AI arms race is a bad idea, and should be prevented by a ban on offensive autonomous weapons beyond meaningful human control." [11]
- A debate about these concerns took place at the Conference on Certain Conventional Weapons in Geneva in 2019. Only a few countries voted for a ban according to international law. USA, Russia, Israel and UK, countries

with the highest ability to develop such weapons, have little interest in such a prohibition.

- In 2018 thousands of Google employees signed an open letter to CEO Sundar Pichai. They requested, that Google should immediately cancel Maven, a joint project with the US military. "We believe that Google should not be in the business of war. Therefore, we ask that Project Maven be cancelled, and that Google draft, publicize and enforce a clear policy stating that neither Google nor its contractors will ever build warfare technology." [12]
- One of the most prominent initiatives about ethical principles of autonomous systems including AI originates from the Institute of Electrical and Electronics Engineers (IEEE). "The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) has the goal to create specifications for certification and marking processes that advance transparency, accountability, and reduction in algorithmic bias in autonomous and intelligent systems. ECPAIS intends to offer a process and define a series of marks by which organizations can seek certifications for their processes around the A/IS products, systems, and services they provide." [14]
- The Association for Computing Machinery (ACM) organised a conference about AI, Ethics and Society, in February 2018 [15].
- OpenAI, a partnership on AI, brings together companies, industry, nonprofit as well as academic organisations. "OpenAI's mission is to ensure that artificial general intelligence benefits all of humanity. We're a team of a hundred people based in San Francisco, California. The OpenAI Charter [17] describes the principles that guide us as we execute on our mission." [16]
- Several companies establish their own ethic codes on AI, like the IBM's Everyday Ethics for Artificial Intelligence, a document which represents the beginning of a conversation defining everyday ethics for AI [18].
- In 2017 in conjunction with the Asilomar conference Asilomar AI principles were developed. They have been launched by the Future of Life Institute, one of the leading initiatives demanding a responsible development of AI [3].
- The cologne center for ethics, rights, economics and social sciences of health (CERES) held a conference about robot ethics in 2015. Researches discussed questions about ethical problems with intelligent service robots

i. e. in healthcare. Human beings could lose their freedom and become addicted to machines. Humans could become supervised objects by powerful monopoles. Finally this could lead to a loss of privacy and self-determination [19].

6 Conclusion

National level initiatives in Europe are uneven. Some develop rules for robots and artificial intelligence and even adopt legislation. Others are yet to deal with the matter [2]. Actually a machine can't be a legal person, despite the fact that robot Sophia has been officially a citizen of Saudi Arabia since 2017. Still there is a lot of legal work to be done.

"Human dignity as the foundation of human rights implies that meaningful human intervention and participation must be possible in matters that concern human beings and their environment." ([2]) Human beings should determine which values technologies may offer and what is morally relevant. "These cannot be left to machines, no matter how powerful they are." "Moral responsibility, in whatever sense, cannot be allocated or shifted to 'autonomous' technology." ([2])

References

[1] https://ethicsinaction.ieee.org/
[2] Statement on Artificial Intelligence, Robotics and Autonomous Systems, European Group on Ethics in Science and New Technologies, https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf
[3] https://futureoflife.org/ai-principles
[4] https://www.pcmag.com/article/366797/which-self-driving-cars-put-in-the-most-fully-autonomous-mil
[5] https://www.theverge.com/2019/3/5/18252423/uber-wont-be-charged-with-fatal-self-driving-crash-says-prosecutor
[6] https://www.quantamagazine.org/why-alphazeros-artificial-intelligence-has-trouble-with-the-real-world-20180221/
[7] https://new.abb.com/news/detail/11164/autonomous-systems
[8] https://cra.org/ccc/wp-content/uploads/sites/2/2016/08/Autonomous-Systems-WG-Overview-final.pdf
[9] https://www.iep.utm.edu/autonomy/

[10] https://www.electronicdesign.com/automotive/ethical-considerations-autonomous-system-design
[11] https://futureoflife.org/open-letter-autonomous-weapons/
[12] https://www.googlewatchblog.de/2018/04/offener-brief-google-ceo/
[13] https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?__blob=publicationFile
[14] https://standards.ieee.org/industry-connections/ec/autonomous-systems.html
[15] http://www.aies-conference.com
[16] https://openai.com/
[17] https://openai.com/charter/
[18] https://www.ibm.com/ibm/responsibility/ibm_humanrightsprinciples.htm
[19] https://www.ceres.uni-koeln.de/veranstaltungen/tagungen/roboterethik/
[20] https://www.bostondynamics.com/

A Road Traffic "Behavioral Analysis": Mathematical Modeling and Numerical Simulation

Jean Chamberlain Chedjou and Kyandoghere Kyamakya

Institute of Smart Systems Technologies, Transportation Informatics Group
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

Abstract: This paper considers the mathematical modeling and numerical simulation of macroscopic traffic flow on a road segment (with finite length) subdivided into three sections. A set of coupled equations is obtained to express the temporal evolution of the fundamental parameters of traffic (i. e., density, speed, and flow) on each of the three sections of the road segment. A numerical simulation of the set of coupled equations is carried out in order to depict the possible states of traffic flow. Three main states of traffic are depicted namely, the state undersaturation, the state at saturation, and the state oversaturation (ending to complete jam). The aforementioned traffic states are depicted through the plots of various 2D-fundamental diagrams expressing the mutual effects of "flow versus density", "speed versus flow", and "speed versus density". In order to validate the results obtained a benchmarking is considered, which consists of comparing the fundamental diagrams obtained (in this paper) with the fundamental diagrams provided by the classical and popular Greenshields models. The comparison shows a good agreement between the results.

1 Introduction

During the past decades several road traffic management policies, strategies, and techniques have been developed to solve inherent traffic problems such as congestion, traffic jams, etc. The addressing of aforementioned problems has necessarily led to time saving, reduction of energy consumption, reduction of pollution and accidents (just to name a few) and thereby has significantly contributed to increasing the quality of life of the society. Despite the tremendous attention devoted to the development of traffic management policies and techniques for traffic analysis and prediction, mathematical models also appear as good candidates for the efficient modeling of the complex dynamical behavior

undergoes by traffic flow in all possible traffic states ranging from the undersaturated to saturated states, and also from the oversaturated to chaotic states. Moreover, mathematical models can efficiently express some complex insights of traffic flow such as shock wave, stop-and-go wave, rarefaction wave, synchronized traffic and chaotic traffic dynamics, just to name a few. A further advantage of using mathematical models is the possibility of controlling the aforementioned states of traffic flow through monitoring of specific parameters/ coefficients of the resulting mathematical models. Mathematical models are also viewed as framework of predilection for the numerical simulation of traffic flow and a quick evaluation of traffic dynamics. This evaluation is essential for the prediction and control of a possible occurrence of aforementioned traffic states and thus/thereby offering the possibility of proactively cancelling or avoiding some unnecessary states of traffic flow. Therefore mathematical models constitute a good and flexible framework that can be potentially considered and used in various contexts such as traffic planning and optimized adaptive and proactive traffic management [1]. Also, mathematical traffic models can be efficiently used to perform various online-simulations of the spatiotemporal and complex dynamics of traffic flows. The online-simulation based mathematical models is advantageous as it satisfactorily addresses key/fundamental issues (e. g. reliability, correctness, robustness, scalability, real-time computing, etc.) characterized by metrics such as accuracy, high computation speed, low memory consumption, efficient control/ optimization and forecasting capability of traffic flows, just to name a few.

The seminal model for traffic flow was proposed by Lighthill-Whitham (1955) and Richards (1956), under the acronym of LWRs model. This model is of first order and is based on the continuity equation from compressible dynamics theory. The LWRs model expresses the conservation of a flowing quantity from one point to another. It is known (from the literature) that the LWR model can efficiently handle the evolution of shockwaves [2, 11]. However the LWR model is subject to limitations as it does assume both constant speed and infinite acceleration of vehicles. In order to tackle the aforementioned drawbacks of the LWR model, the Payne model was developed. However, it was further demonstrated that the Payne model does not integrate/respect the anisotropic principle. This drawback could lead to negative values of speeds when using the Payne model. This latter specific drawback was further tackled through the development of a class of models such as Zhang [3], Jiang et al. [4], and Gupta and Katiyar [5], just to name a few.

In this paper, we develop and simulate mathematical models of traffic flow, especially using Partial Differential Equations (PDEs), which are further expressed into coupled discrete models for the sake of numerical computation. The coupled mathematical models are obtained in case of a road segment subdivided into three sections. The resulting models are solved numerically and the numerical findings are exploited to demonstrate the coexistence between several different traffic states (e. g. the state under saturation, the state at saturation and the state over saturation). The aforementioned traffic states are clearly depicted through the plot of various 2D- fundamental diagrams showing the interaction (or mutual effects) between the fundamental parameters of traffic flow at macroscopic level of detail namely, the average speed, average flow and average density.

The paper is organized as follows. The second part is concerned with the mathematical modeling. Mathematical models are derived in order to express the traffic flow on a road segment of finite length subdivided into three sections. The third part is devoted to numerical simulation. The resulting mathematical models are solved numerically in order to depict the striking dynamical behavior of traffic flow on the road segment at stake. The results obtained numerically are further compared with results provided by the classical Greenshields models. Finally, some concluding remarks are formulated in part 4 along with some pending unsolved research issues as outlooks.

2 Mathematical Modeling

In order to facilitate the modeling procedure of the traffic flow on a road segment of finite length, we consider the general representation in figure 1. This figure shows a multiclass traffic flow encompassing different types of vehicles. The traffic depicted in figure 1 is a multi-lanes traffic flow with overtaking possibility. Also, the traffic flow in figure 1 is subject to external disturbances due to the presence of both on-ramps (entrance ramps) and off-ramps (exit ramps). In figure 2 we restrict our analysis to a single section of the road segment. As it appears in figure 2, the single section of a road segment is considered as a dynamical system with one input $q_{in}(t)$, one output $q_{out}(t)$ and two external disturbances resulting from on-ramp $r(t)$ and off-ramp $s(t)$. The quantity $k(t)$ (in figure 2) expresses the temporal evolution of traffic flow on the section of road segment shown in figure 2. Therefore expressing the relationship between all the parameters defined above leads to a fundamental analytical relationship representing the mathematical model of traffic flow. For the sake of generaliza-

tion, the mathematical models obtained when considering a single section of a road segment can be used to deduce the models corresponding to the case of a road segment subdivided into three and six sections. This latter and unsolved research idea does not fall within the scope of this paper and is currently being explored as an outlook.

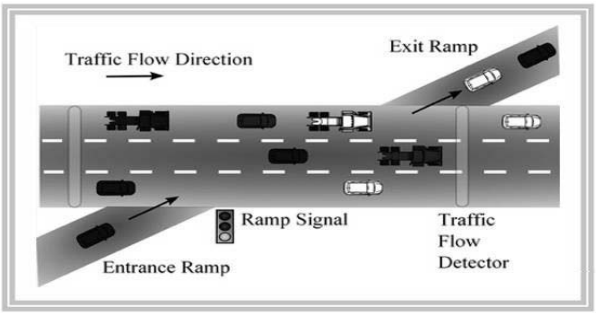


Fig. 1: General representation of the traffic flow on a road segment of finite length. This representation shows three lanes with overtaking possibility. Both on-ramp and off-ramp are also illustrated [6].

We now consider the modeling of traffic flow on a road segment subdivided into three sections of different lengths as shown in figure 3. Each of the three sections in figure 3 is made-up of one input and one output. Therefore the first step of the modeling procedure consists of evaluating the total incoming traffic flow at each section $S_i(i = 1, 2, 3)$; this represents the input traffic flow of the section. We also evaluate the total outgoing traffic flow from each section; this represents the output traffic flow of the section.

Let us denote the spatiotemporal evolution of the traffic density on each section of the road by ρ_i, the spatiotemporal evolution of the speed of cars/vehicles on each section of the road by v_i and the spatiotemporal evolution of the traffic flow on each section of the road by $q_i = \rho_i v_i$. The upstream flow q_0 and the upstream speed v_0 stand respectively for the initial flow and the initial speed of traffic as shown in figure 3. The parameter α is the weight factor. This factor determines the relative impact/influence of the average traffic volume on a given section of the road segment (say, S_1) on the average traffic volume on the other section of the road segment (say, S_2). Therefore α denotes the strength of the

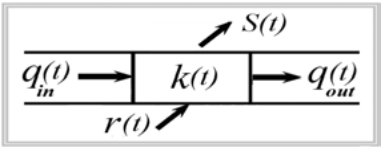


Fig. 2: Representation of traffic flow in a single section of the road segment. This representation considers a single section of the road segment as a dynamical system with inputs and outputs. The simple representation here is considered for the sake of facilitating the modeling procedure.

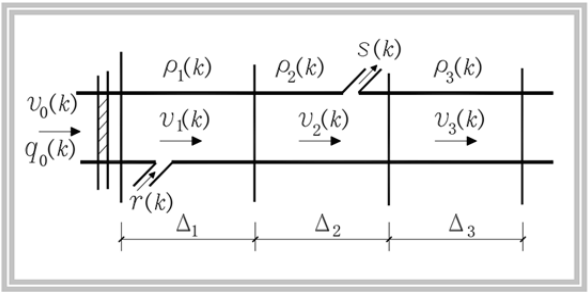


Fig. 3: Representation of traffic flow on a road segment subdivided into three sections of different lengths.

direct coupling between neighboring sections of the road segment subdivided into multiple sections of different lengths.

According to the description provided to spatiotemporal variables ρ_i, v_i and q_i and according to the definition provided to the parameter α, the inputs and outputs of each of the three sections of the road segment are expressed as follows:

$$Section\, S_1: \quad \begin{cases} Input\,1: q_{in} &= [q_0 + r(t)] \\ Output\,1: q_{out} &= [\alpha\, \rho_1\, v_1 + (1-\alpha)\, \rho_2\, v_2] \end{cases} \tag{1}$$

$$Section\, S_2: \quad \begin{cases} Input\,2: q_{in} &= [\alpha\, \rho_1\, v_1 + (1-\alpha)\, \rho_2\, v_2] \\ Output\,2: q_{out} &= [\alpha\, \rho_2\, v_2 + (1-\alpha)\, \rho_3\, v_3 + s] \end{cases} \tag{2}$$

$$Section\, S_3: \quad \begin{cases} Input\,3: q_{in} &= [\alpha\, \rho_2\, v_2 + (1-\alpha)\, \rho_3\, v_3] \\ Output\,3: q_{out} &= [\alpha\, \rho_3\, v_3 + (1-\alpha)\, \rho_3^*\, v_3^*] \end{cases} \tag{3}$$

In order to combine the inputs and outputs expressed in equations (1) – (3) we use the continuum form of the LWRs model expressed in (4) [7, 11].

$$\frac{\partial \rho}{\partial t} + \frac{\partial (q)}{\partial x} = 0 \tag{4}$$

The continuum form of the LWRs model in (4) can be transformed into discrete form as follows:

$$\rho(n+1) = \rho(n) + \frac{\Delta t}{\Delta x}\left[q_{in}(n) - q_{out}(n)\right] \tag{5}$$

In (5) the quantities Δt and Δx stand for the temporal step size and the spatial step size respectively. In order to preserve the accuracy of the discrete form (5), the quantities Δt and Δx must be chosen very small ($\Delta t \to 0$ and $\Delta t \to 0$) to make sure that the discrete form (5) can accurately approximate the continuous form (4). However the above condition for preserving the accuracy of the discrete form (5) is a purely theoretical mathematical condition which need to be adapted to practical constraints of the mathematical modeling procedure carried out in this part/section.

A typical illustration of how the values of parameters Δt and Δx of the discrete form (5) can be chosen under some practical constraints is as follows: *Constraint 1.* According to the spatial discretization performed in (5), Δx represents the spatial distance between the system input (see q_{in} in (5)) and the system output (see q_{out} in (5)). Therefore the parameter Δx in (5) represents the length of each of the three sections of the road segment in figure 3. *Constraint 2.* According to the temporal discretization performed in (5), Δt represents a temporal duration (i. e., a time slot or a time period) in which the variation of each fundamental parameter of traffic flow at macroscopic level of detail (say, variation of "average flow", variation of "average speed" and variation of "average density") can be measured. These two key constraints are exploited here to combine the discrete form (5) with Eqs. (1) – (3).

Equation (5) is the key fundamental expression which is now applied to the set of Equations (1) – (3) to obtain the mathematical model of traffic flow on a road segment subdivided into three sections (see figure 3).

Combining (5) and (1) leads to the mathematical model (see (6)) of traffic flow on the first section of the road segment in Fig. 3.

$$\rho_1(n+1) = \rho_1(n) + \frac{\Delta t}{\Delta x}\left[q_0 - \alpha\,\rho_1(n)\,v_1(n) - (1-\alpha)\,\rho_2(n)\,v_2(n) + r\right] \tag{6}$$

Similarly, (5) can be combined with (2) to obtain the mathematical model of traffic flow (see (7)) on the second section of the road segment in Fig. 3.

$$\begin{aligned}\rho_2(n+1) &= \rho_2(n) + \tfrac{\Delta t}{\Delta x}\left[\alpha\,\rho_1(n)\,v_1(n) + (1-\alpha)\,\rho_2(n)\,v_2(n) - \right.\\ &\quad \left. \alpha\,\rho_2(n)\,v_2(n) - (1-\alpha)\,\rho_3(n)\,v_3(n) - s\right]\end{aligned} \tag{7}$$

The discrete form (5) is further combined with (3) to obtain the mathematical model of traffic flow (see (8)) on the third section of the road segment in Fig. 3.

$$\begin{aligned}\rho_3(n+1) &= \rho_3(n) + \tfrac{\Delta t}{\Delta x}\left[\alpha\,\rho_2(n)\,v_2(n) + (1-\alpha)\,\rho_3(n)\,v_3(n) - \right.\\ &\quad \left. \alpha\,\rho_3(n)\,v_3(n) - (1-\alpha)\,\rho_3^*(n)\,v_3^*(n)\right]\end{aligned} \tag{8}$$

In (8) the quantities ρ_3^* and v_3^* are defined as follows: $\rho_3^*(n) = \rho_3(n-1)$ and $v_3^*(n) = v_3(n-1)$.

In the discrete expressions (6) – (8), six variables are clearly identified, namely the densities (ρ_1, ρ_2, and ρ_3) and the speeds (v_1, v_2, and v_3). Therefore, the solving of equations (6) – (8) requires three additional relationships between the aforementioned variables. These relationships correspond to equations for the evolution of speeds (v_1, v_2, and v_3), obtained from the basic METANET model [8] (see also in Ref. [9]). According to [8] and [9], the evolution of speeds is modeled by the three additional relationships in (9) – (11).

$$\begin{aligned}v_1(n+1) &= v_1(n) + \frac{\Delta t}{\tau}\left[V(\rho) - v_1(n)\right] + \frac{\Delta t}{\Delta x} v_1(n)\left[v_0 - v_1(n)\right] + \\ &\quad \frac{\Delta t}{\Delta x}\frac{\gamma}{\tau}\left[\frac{\rho_2(n) - \rho_1(n)}{\rho_1(n) + \lambda}\right]\end{aligned} \tag{9}$$

$$\begin{aligned}v_2(n+1) &= v_2(n) + \frac{\Delta t}{\tau}\left[V(\rho) - v_2(n)\right] + \frac{\Delta t}{\Delta x} v_2(n)\left[v_1 - v_2(n)\right] + \\ &\quad \frac{\Delta t}{\Delta x}\frac{\gamma}{\tau}\left[\frac{\rho_3(n) - \rho_2(n)}{\rho_2(n) + \lambda}\right]\end{aligned} \tag{10}$$

$$\begin{aligned}v_3(n+1) &= v_3(n) + \frac{\Delta t}{\tau}\left[V(\rho) - v_3(n)\right] + \frac{\Delta t}{\Delta x} v_3(n)\left[v_2 - v_3(n)\right] + \\ &\quad \frac{\Delta t}{\Delta x}\frac{\gamma}{\tau}\left[\frac{\rho_3(n-1) - \rho_3(n)}{\rho_3(n) + \lambda}\right]\end{aligned} \tag{11}$$

Equations (9) – (11) contain, each, three main terms. The first term corresponds to the relaxation term [8]. This term expresses the wish (by the drivers) of achieving the desired speed denoted by $V(\rho)$ for the actual/current traffic density

denoted by ρ. The second term corresponds to the convection term [8]. This term expresses the change in speed due to the inflow of vehicles with a different speed. The third term corresponds to the anticipation term [8]. This term expresses the change of speeds by drivers when the downstream density changes.

Finally the set of coupled equations (6) – (11) constitute the mathematical model of traffic flow on the road segment (in Fig. 3) subdivided into three sections. The densities (ρ_1, ρ_2, and ρ_3) and speeds (v_1, v_2, and v_3) are obtained as direct numerical simulation of (6) – (11). The corresponding traffic flows (q_1, q_2, and q_3) are deduced from the fundamental relationship between Speed-Flow-Density (i. e., average flow $q_1 = \rho_1 v_1$, average flow $q_2 = \rho_2 v_2$, and average flow $q_3 = \rho_3 v_3$).

3 Numerical Simulation

The numerical simulation aims at validating the mathematical model obtained (and expressed by the set of coupled equations (6) – (11)). Therefore a benchmarking is conducted aiming at comparing the results obtained through the numerical simulation of the set of coupled equations (6) – (11) with the results provided by the classical Greenshields fundamental diagrams [10]. These diagrams are obtained as results of the plot of the Greenshields models in (12) – (14) (See Ref. [10]). In (12) – (14), the flow is denoted by y_i, the density is denoted by x_i and the speed is denoted by z_i The 2D-fundamental diagram "flow versus density" is obtained from (12) while the 2D-fundamental diagram "speed versus flow" is obtained from (13). Also, the 2D-fundamental diagram "speed versus density" is obtained from equation (14).

$$y_i = \left[v_f \, x_i - \frac{v_f}{K_{Ji}} \, x_i^2 \right] \tag{12}$$

$$z_i^2 = \left[v_f \, z_i - \frac{v_f}{K_{Ji}} \, y_i \right] \tag{13}$$

$$z_i = \left[v_f \; - \frac{v_f}{K_{Ji}} \, x_i \right] \tag{14}$$

In (12) – (14), i is the index of the section ($i = 1, 2, 3$) of the road segment in figure 3, v_f represents the free flow speed and K_{Ji} is the jam density in a section with index i.

3.1 Numerical Simulation of the PDE Model

We now want to plot the evolution of "flow" versus "density", "speed"versus "flow", and "speed"versus "density" in three sections of the road segment. The fundamental parameters of traffic (i. e., "density", "speed" and "flow") are obtained as numerical solution of the mathematical model of traffic flow in (6) – (11). The values of parameters used for numerical simulation are defined as follows: $r = 0$, $s = 0$, $v_0 = 90\,\text{km/h}$, $q_0 = 3000\,\text{V/h}$, $T = 9\,\text{s}$, $D = 0.800\,\text{km}$, $\alpha = 0.95$, $\tau = 10\,\text{s}$, $\gamma = 34.7$, $\lambda = 40$, $v_f = 110\,\text{km/h}$, $b = 2.08$, $X_{cr} = 31$, and $K_{Ji} = (4 * q_{max}(i))/v_f$. Note that $q_{max}(i)$ stands for the road capacity in each of the three sections with index "i" ($i = 1, 2, 3$). The maximum duration of simulation is $T_{max} = 5\,\text{hours}$ and the maximum number of iterations is $m_{max} = 2000$; this corresponds to the temporal step size of $T = 9\,\text{s}$. The aforementioned values of parameters are fixed (i. e., constant values) and are used to obtain the results of numerical simulation presented below.

The first analysis consists of using the values of parameters defined above to solve the coupled equations (6) – (11). The numerical results obtained are depicted in figures 4 a), 4 b), and 4 c). These results (which are labelled as PDE model) clearly reveal the under-saturated state of traffic flow as well as the saturated and over-saturated states of traffic flow. In contrast, the state of jam traffic is not detected by the PDE model. Another interesting comment, which could be formulated based on results provided by the PDE model (equivalently by the set of coupled equations (6) – (11)) is that the capacity of the first section of the road segment is approximately equal to $q_{max}(1) = 8115\,\text{Vehicles/hour}$. Considering sections 2 and 3, the respective capacities are approximately equal to $q_{max}(2) = 11970\,\text{Vehicles/hour}$ and $q_{max}(3) = 14910\,\text{Vehicles/hour}$. The traffic densities at capacities are approximately equal to $x_1 = 137\,\text{Vehicles/km}$ (in section 1), $x_2 = 225\,\text{Vehicles/km}$ (in section 2), and $x_3 = 295\,\text{Vehicles/km}$ (in section 3).

The second analysis is focused on the evolution of "speed" versus "flow" in three sections of the road segment in figure 3. The numerical results obtained are depicted in figures 5 a), 5 b), and 5 c). These results (which are labelled as PDE model) clearly reveal the existence of the state under-saturation in all three sections of the road segment. This state corresponds to traffic flows (or traffic volumes) less than 3000 Vehicles/hour. Beyond the threshold of 3000 Vehicles/hour, the states of saturation and oversaturation are observed. The speeds at capacity are approximately $x_4 = 59\,\text{km/h}$ (in section 1), $x_5 = 53\,\text{km/h}$ (in section 2), and $x_6 = 50\,\text{km/h}$. (in section 3). The capacities measured in figures 5 a) –

5 c) are 8112 Vehicles/hour (in section 1), 11969 Vehicles/hour (in section 2) and 14914 Vehicles/hour (in section 3). These results are in good agreement with the results obtained in figures 4 a), 4 b), and 4 c).

The third analysis is concerned with the evolution of "speed" versus "density". The numerical solution of the coupled equations (6) – (11) is used for the analysis. The results obtained are depicted in figures 6 a), 6 b) and 6 c). As it appears in figures 6 a) – 6 c), the evolution of "speed" versus "density" is linear at low traffic density. In contrast this evolution is nonlinear at high traffic density. As illustrative example, if we consider the traffic flow in section 1, the evolution of "speed" versus "density" is almost linear when the traffic density is below the threshold of $x_1 = 190$ Vehicles/km. Beyond this threshold the evolution is nonlinear. Similarly, in section 2, the evolution of "flow" versus "density" is linear below the threshold of $x_2 = 320$ Vehicles/km. Beyond this threshold the evolution is nonlinear. Considering section 3, the evolution of "flow" versus "density" is linear below the threshold of $x_3 = 455$ Vehicles/km and the evolution is nonlinear beyond the threshold.

3.2 Numerical Simulation of the Greenshields Models

The values of parameters used in the preceding case (i. e., case of the PDE-model) are now used to solve the Greenshields models (12) – (14). The results obtained are commented below.

The results obtained as numerical solutions of (12) are depicted in figures 4 a), 4 b), and 4 c) (see dots labelled as Greenshields model). These figures show the evolution of "flow" versus "density" in all three sections of the road segment. Figure 4 a) corresponds to the evolution of "flow" versus "density" in section 1 ($i = 1$), figure 4 b) corresponds to "flow" versus "density" in section 2 ($i = 2$) and figure 4 c) corresponds to "flow" versus "density" in section 3 ($i = 3$). Comparing the results obtained in figures 4 a) – 4 c) using the PDE-model with results obtained in figures 4 a) – 4 c) using the Greenshields model, a good agreement is obtained when the traffic flow is in the state under-saturation. However, a divergence is observed between the two methods/models when the traffic flow is in the state over-saturation. The rate of divergence in section 1 is greater than the rate in section 2. Also, the rate of divergence in section 2 is greater than the rate in section 3. Consequently, it appears (from figures 4 a) – 4 c)) that a good agreement is observed between the two methods/models in the oversaturated state of traffic only in the third section of the road.

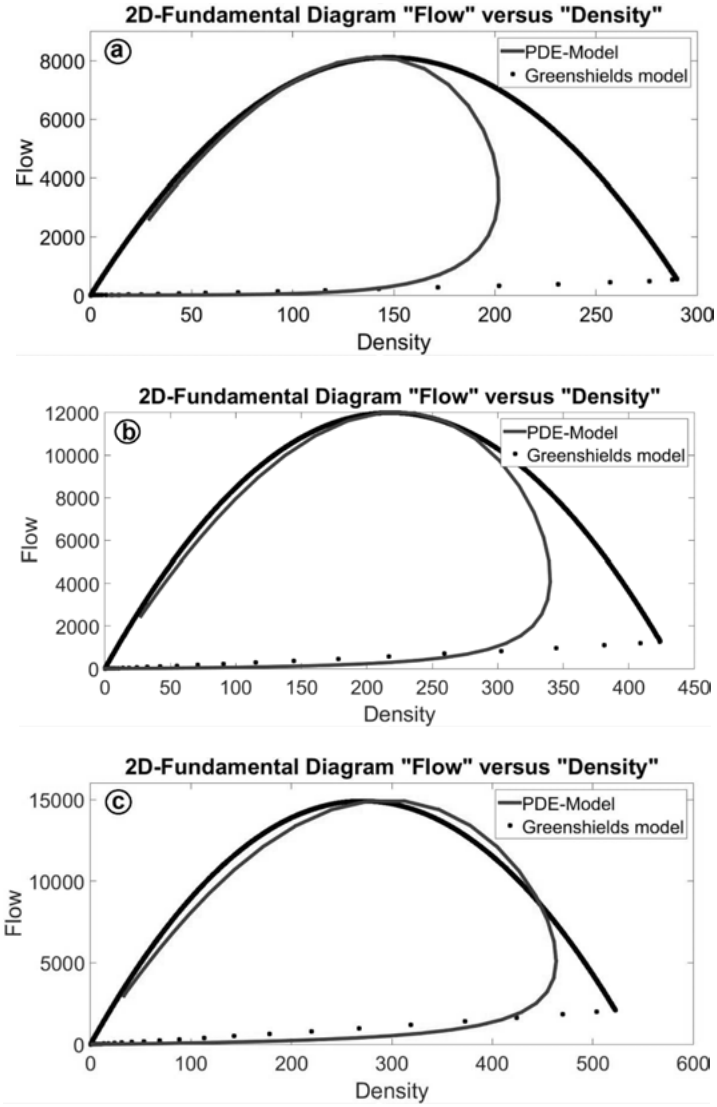


Fig. 4: Fundamental diagram expressing the evolution of "flow" versus "density" in the three sections of the road segment in figure 3. The plots in (a), (b), and (c) correspond to sections 1, 2, and 3, respectively.

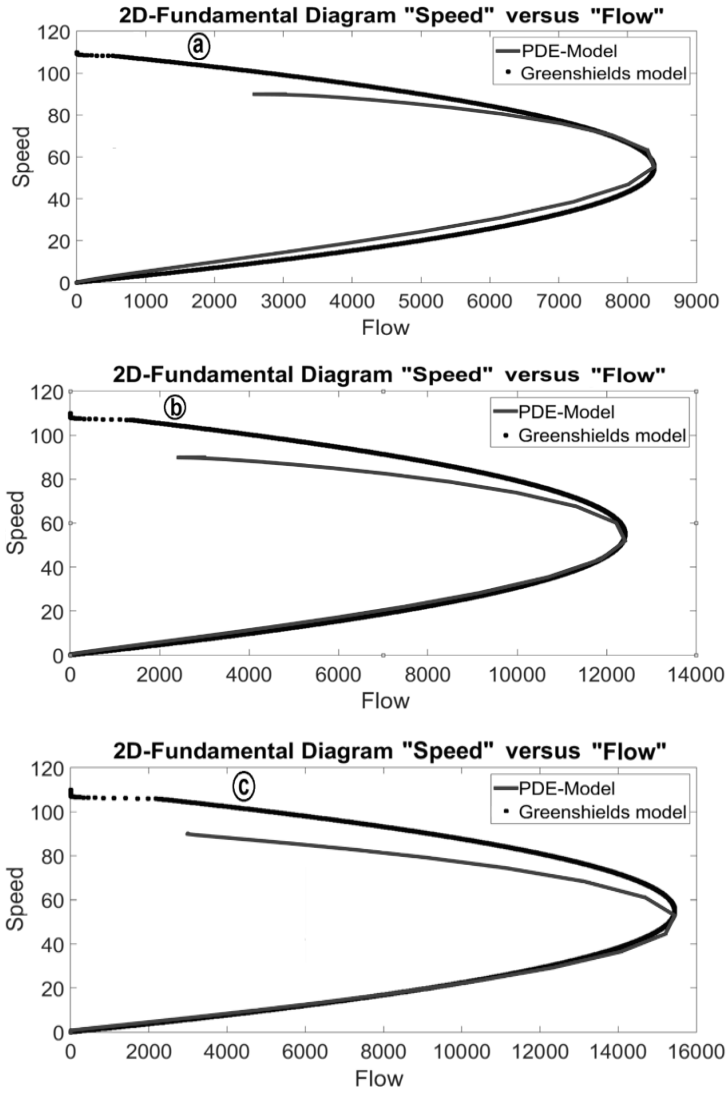


Fig. 5: Fundamental diagram expressing the evolution of "speed" versus "flow" in the three sections of the road segment in figure 3. The plots in (a), (b), and (c) correspond to sections 1, 2, and 3, respectively.

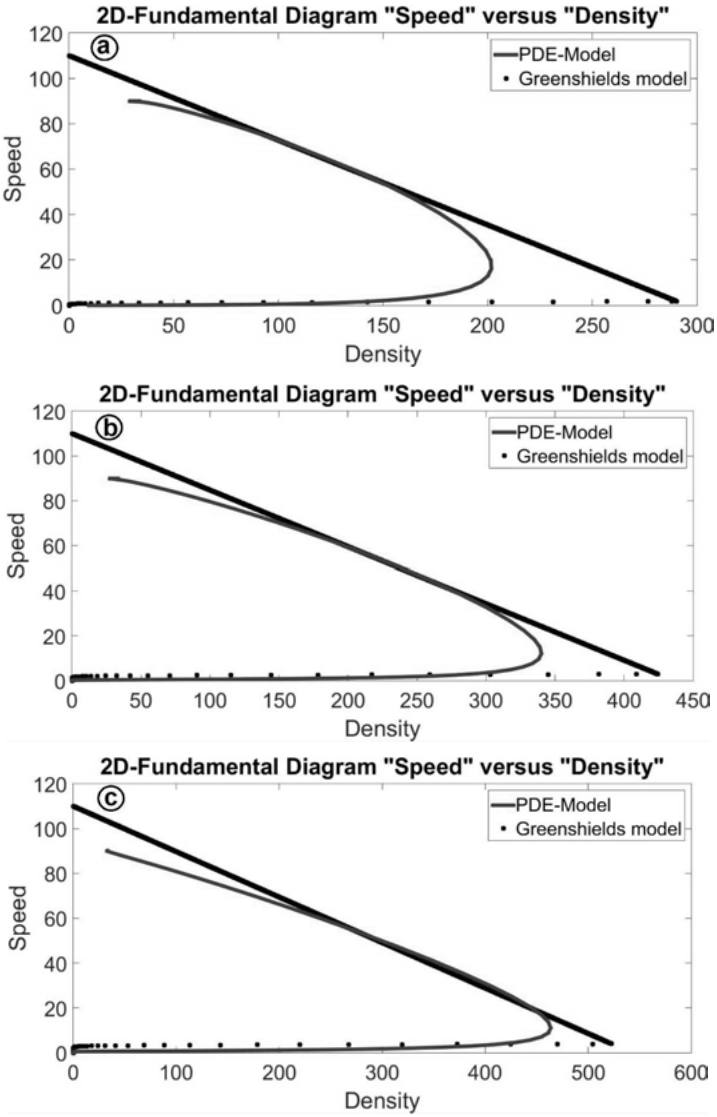


Fig. 6: Fundamental diagram expressing the evolution of "speed" versus "density" in the three sections of the road segment in figure 3. The plots in (a), (b), and (c) correspond to sections 1, 2, and 3, respectively.

The results obtained as numerical solutions of (13) are depicted in figures 5 a), 5 b), and 5 c) (see dots labelled as Greenshields model). These figures show the evolution of "speed" versus "flow" in all three sections of the road segment. Figure 5 a) corresponds to the evolution of "speed" versus "flow" in section 1 ($i = 1$), figure 5 b) corresponds to "speed" versus "flow" in section 2 ($i = 2$) and figure 5 c) corresponds to "speed" versus "flow" in section 3 ($i = 3$). Comparing the results obtained in figures 5 a) – 5 c) using the PDE-model with results obtained in figures 5 a) – 5 c) using the Greenshields model, a good agreement is obtained when the traffic flow is in the state under-saturation. However, a divergence is observed between the two methods/models when the traffic flow is in the state over-saturation. The rate of divergence in section 1 is less than the rate in section 2. Also, the rate of divergence in section 2 is less than the rate in section 3. Consequently, it appears (from figures 5 a) – 5 c)) that a good agreement is observed between the two methods/models in the oversaturated state of traffic only in the first section of the road segment.

Finally, the results obtained as numerical solutions of (14) are depicted in figures 6 a), 6 b), and 6 c) (see dots labelled as Greenshields model). These figures show the evolution of "speed" versus "density" in all three sections of the road segment. Figure 6 a) corresponds to the evolution of "speed" versus "density" in section 1 ($i = 1$), figure 6 b) corresponds to "speed" versus "density" in section 2 ($i = 2$) and figure 6 c) corresponds to "speed" versus "density" in section 3 ($i = 3$). Comparing the results obtained in figures 6 a) – 6 c) using the PDE-model with results obtained in figures 6 a) – 6 c) using the Greenshields model, a good agreement is obtained when the traffic flow is in the state under-saturation. However, a divergence is observed between the two methods/models when the traffic flow is in the state over-saturation. The rate of divergence in section 1 is greater than the rate in section 2. Also, the rate of divergence in section 2 is greater than the rate in section 3. Consequently, it appears (from figures 6 a) – 6 c)) that a good agreement is observed between the two methods/models in the oversaturated state of traffic only in the third section of the road.

4 Conclusion and Outlook

In this work, we have considered the mathematical modeling and numerical simulation of traffic flow on a road segment subdivided into three sections. A mathematical model has been developed to express the dynamics of traffic. The mathematical model developed was further solved numerically and the solutions obtained were used to analyze the dynamics of traffic in the three sections

of the road segment. The outcome of numerical simulation has revealed three main interesting states of traffic namely, the undersaturated, saturated and oversaturated states. In order to validate the mathematical model developed, a benchmarking has been considered. The benchmark has consisted of comparing the results obtained using the method/model developed (in this work) with results obtained using the classical Greenshields model. A quick qualitative comparison of the plots obtained has revealed a good agreement between method 1 (i. e., the model developed in (6) – (11)) and method 2 (i. e., the Greenshields model in (12) – (14)) when the traffic is in the state undersaturation, and also in the state at saturation. However it has been observed that when the traffic states evolve from the state oversaturation to jam/congestion, a significant divergence appeared between the results provided by the two methods.

Let us note that the comparison performed in this work was essentially qualitative and nor quantitative. Therefore it could be nicer to define a quantitative metrics (e. g. mean squared error (MSE) also called mean squared deviation (MSD)), which could be used to compare all results provided by the two methods/models. Another interesting issue currently under investigation is the analysis of stability of the mathematical model developed in this work. This analysis could lead to the discovery of equilibrium points/states which corresponds to equilibrium traffic in practice. The bifurcation analysis of the dynamics of traffic flow is also currently under investigation based on the mathematical model developed in this work. The upstream traffic flow is used as control parameter of the bifurcation analysis. The main aim of the bifurcation analysis currently under investigation is to demonstrate the existence of chaos in the road scenario envisaged in this work. Therefore the ranges/windows of the control parameter within which the chaotic dynamics is observed could be proactively exploited to control the traffic dynamics and separate the stable states of traffic from the unstable and/or chaotic traffic states. The aforementioned unsolved issues are currently under investigation as outlooks.

References

[1] Allan M de Souza et al., "Traffic management systems: A classification, review, challenges, and future perspectives", International Journal of Distributed Sensor Networks (IJDSN), Vol. 13, issue: 4, pp. 1–14, 2017

[2] Wen-Long Jin, "A Multi-commodity Lighthill-Whitham-Richards Model of Lane-changing Traffic Flow", Transportation Research Part B: Methodological, vol. 57, pp. 361–377, 2013.

[3] H.M. Zhang, "New Perspectives on Continuum Traffic Flow Models", Networks and Spatial Economics, vol. 1, no. 1, pp. 9–33, 2001.

[4] Rui Jiang, Qing-Song Wu , Zuo-Jin Zhu, "A new continuum model for traffic flow and numerical tests", Transportation Research Part B: Methodological, vol. 36, no. 5, pp. 405–419, 2002.

[5] A. K. Gupta and V.K. Katiyar, "A new anisotropic continuum model for traffic flow", Physica A, vol. 368, pp. 551–559, 2006.

[6] N. Shlayan and P.Kachroo, "Feedback Ramp Metering using godunov method based hybrid model", J.Dyn.Sys., Meas and Control, vol. 135, no. 15, 2013.

[7] N. Polson and V. Sokolov, "Bayesian analysis of traffic flow on interstate I-55: The LWR model", The Annals of Applied Statistics, vol. 9, No. 4, 1864–1888, 2015.

[8] M. van den Berg, A. Hegyi, B. De Schutter, and J. Hellendoorn, "Integrated traffic control for mixed urban and freeway networks: A model predictive control approach", European Journal of Transport and Infrastructure Research, vol. 7, no. 3, pp. 223–250, Sept. 2007.

[9] Daijian Ling, and Peng Xiao, "Simulation Research on Chaos Characteristics of a Class of Macroscopic Traffic Models", in: *International Conference on Engineering Computation*, pp. 135–138, 2009.

[10] Mashrur A. Chowdhury and Adel Sadek, "Fundamentals of Intelligent Transportation Systems Planning", Artech House, ISBN: 9781580531603, 2003.

[11] Durgesh Vikram, Sanjay Mittal, and Partha Chakroborty, "Exploring the Behavior of LWR Continuum Models of Traffic Flow in Presence of Shock Waves", Procedia – Social and Behavioral Sciences, vol. 104, 2013.

Natural Language Processing

Dynamic Clustering for Segregation of Co-occurrence Graphs

Supaporn Simcharoen and Herwig Unger

Chair of Communication Networks, FernUniversität in Hagen, Germany

Abstract: Clustering in co-occurrence graphs – similar to knowledge in the humans' brain – are increasing while reading one document after the other. Each sentence is read, stemming and stopwords removal are processed successfully, new words and relations between words are added, a co-occurrence graph is built. Then each word will be assigned to a cluster. By using the distance from the cluster center to a node, this ensures that the node is a member of that cluster. The cluster center (Centroid) is a node or a word that has the shortest average distance to every other node in the graph. After the clustering is stable, all centroids are read to determine hierarchy, then the inter-clusters are opened. The next hierarchy level is built until the last node (root) remains. Experimental results will be later discussed.

1 Introduction and Motivation

Clustering is helpful in grouping, decision making and machine learning situations including data mining, document retrieval, image segmentation, and pattern classification [1]. Clustering is an effective method in search computing [2]. It allows to group similar results [3]. Data clustering is an effective unsupervised data mining technique used to discover knowledge within the data [4]. Text categorization is the process that can manage electronic documents effectively [5]. The objects are sorted into categories, usually for some specific purpose. Text classification is the automatic determination of the text relation according to the real content under the given category topic [6].

Clustering is used in several research communities to describe processes for grouping of unlabeled data. Clustering is a useful and difficult task, the modular segregation process found in the brain during a learning process [7] is an idea that can solve this challenge.

As a child, the learning process in the humans' brain begins to learn a word and the number of words increase steadily. When children grow up, the relations

between words is learned. After learning more words with more connections, some words and some connections become stronger. Then those words will be organized into various categories to form a repeated learning. Our brains begin to learn that father, mother, and kids; and understand that a category of these words is family. Similarly, rice, vegetables, shrimp, and fish; are from a category of food. These categories will be even more clear and increase gradually when an adult grows up, see Fig. 1.

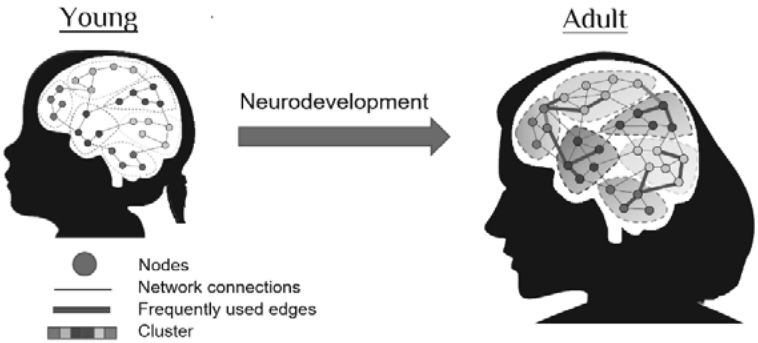
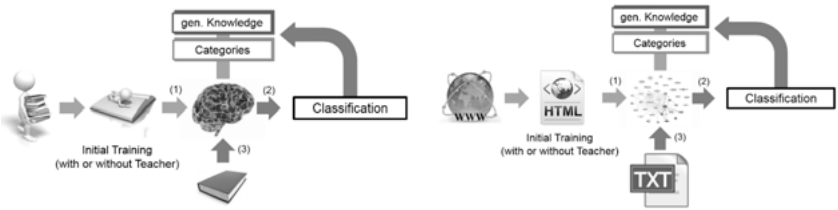


Fig. 1: Modular segregation process in the brain while learning

Fig. 1 is inspired by a neuroscience publication [7] on modular segregation which shows inspiring comparison. Learning within the human brain starts with reading a text from documents or books (1). Then the brain begins to learn new words, building categories for this knowledge and sorting them into the categories (2) in order to be able to classify other categories later. Moreover, adding unknown texts into them usually based on some similarity considerations (3), see Fig. 2a.

Therefore, four processes appear in the brain while reading a text from documents or books (following [7] and Fig. 1).

1. New words are learned.
2. Relations (co-occurrences) between word are added and networks (e. g. co-occurrence graphs) appear.
3. Clusters consequently occur in these networks.
4. Texts are categorized, depending on recent state of the networks.



(a) Process in the brain (b) Process in the model

Fig. 2: The Processes while reading

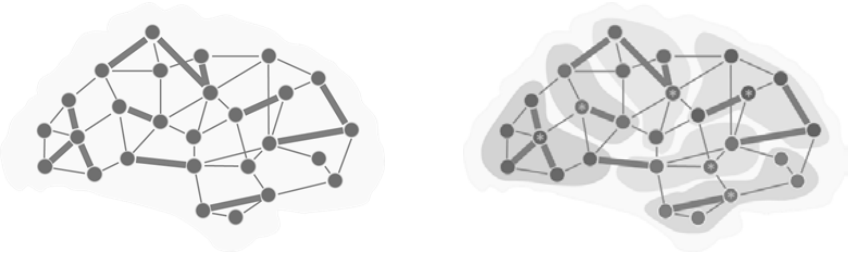
To formalize these processes in a model by replacing the storage of the brain to a co-occurrence graph, which growing while reading one document after the other, words are nodded, relations are edged, and clusters are clustered in the co-occurrence graph. Finally, a co-occurrence graph is consequently used for document categorization, see Fig. 2b.

2 Cluster Building

2.1 Concept

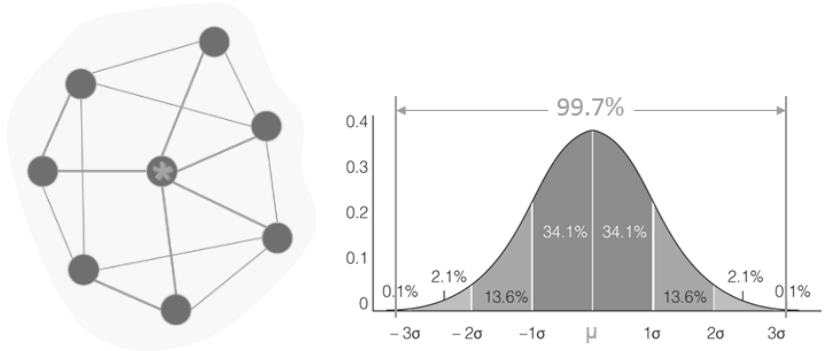
While one document after the other is read, sentence by sentence, each word of each sentence and relation (distance) between words are added. Then, a co-occurrence graph is built, see Fig. 3a. At the same time that each word which is added to the clustering process become active, each word of the cluster will be assigned by depending on the distance to *the cluster center (centroid)*. Then the clusters in a co-occurrence graph is built, see Fig. 3b (a node with (*) is the cluster center). Their significance values have been determined by using the Dice coefficient [8]. An embedded Neo4j (`https://neo4j.com`) is used to save all words (represent the terms), the relation between words, (represent the co-occurrence, and their significances).

The cluster center (centroid) is a node or a word that has the shortest average distance to every other node in the graph, see Fig. 4a. A cluster center is generated when a new cluster is opened, it is updated when the cluster members increased, and it is used to calculate an average distance (μ) of all nodes from the cluster center. Then, a standard deviation (σ) of distances can be obtained in each cluster.



(a) The co-occurence graph (b) Clusters in the co-occurence graph

Fig. 3: The clustering in the co-occurrence graph



(a) A position of the cluster center (*) (b) The normal distribution (Bell Curve)

Fig. 4: The Cluster center and the normal distribution

As the cluster center works in a discrete environment of a co-occurrence graph, therefore, the distance to the cluster center is used to ensure that the node within distance range ($\mu + 3\,\sigma$) from the cluster center shall be a member of the respective cluster (99.7 % of words are within three standard deviations of the average distance [9], see Fig. 4b).

In each insertion of a new word or change of an edge weight, the evaluation of the clustering change. Clusters can be joined, divided, and restructured. Furthermore, when clusters in the co-occurrence graphs are stable, providing that no more change in the given documents, the clustering will start to create the hierarchy.

2.2 Co-occurence Graph Building Mechanism

The co-occurrence graph [10] starts from one document D_j after the other is read sentence $S_{j,l}$ by sentence, then each word $w_{j,l,i}$ of each sentence and relation (distance d) between words are added. Finally, a co-occurrence graph is built.

Therefore, some basic definitions are needed.

If C be a text corpus which consist of $m = |C|$ documents D_j

$$C = \{D_1 \dots D_m\} \quad .$$

Then every document D_j consist of $k = |D_j|$ sentences $S_{j,l}$, i. e.

$$D_j = \{S_{j,1} \dots S_{j,k}\} \quad .$$

Every sentence $S_{j,l}$ is a composition of $h = |S_{j,l}|$ well-ordered words $w_{j,l,i}$ after stemming and stopwords removal have been applied, i. e.

$$S_{j,l} = \{w_{j,l,1} \dots w_{j,l,h}\} \quad .$$

While reading the word $w_{j,l,i}$ of the sentence $S_{j,l}$ of a document D_j from a corpus, a co-occurrence graph $G(t) = (W(t), E(t), g(E,t))$ is built in a successive manner following the reading steps t, when $t = 0, 1, \dots, \infty$.

The reading for building the co-occurrence graphs consists of the following steps:

1. Start with the initialization:

   a) Set $t = 0$, $W(0) = \emptyset$ and $E(0) = \emptyset$, i. e. start with an empty co-occurrence graph.

   b) Set $j = 1$, $l = 1$ and $i = 1$, i. e. start to read with the first word of the first sentence from the first document.

2. Read $w_{j,l,i}$

3. If $w_{j,l,i} \notin W(t+1)$ set $W(t) := W(t) \cup w_{j,l,i}$ else $W(t+1) = W(t)$

4. If $(i > 1) \wedge \Big((w_{j,l,i}, w_{j,l,h}) \notin E(t)\Big)$ set $E(t) := E(t) \cup (w_{j,l,i}, w_{j,l,h})$
   $\forall\, h, h = 1 \ldots (i-1)$ with $g\Big((w_{j,l,i}, w_{j,l,h}), (t+1)\Big) = 1$
   otherwise set $g\Big((w_{j,l,i}, w_{j,l,h}), (t+1)\Big) = g\Big((w_{j,l,i}, w_{j,l,h}), (t)\Big) + 1$.
   Note that for any two words $w_{a,b,c} = w_{d,e,f}$ only one node is inserted in W for any t.
5. Calculate $\xi^i_{D_j}$ using $G(t)$ following the algorithm given in [10, 11] to assign the document (link) to the respective node.
6. Increase $i := i + 1$
7. If $i < h$ goto 2
8. Set $i = 1$ and increase $l := l + 1$
9. If $l < k$ goto 2
10. Set $l = 1$ and increase $j := j + 1$
11. If $j < m$ goto 2
12. STOP

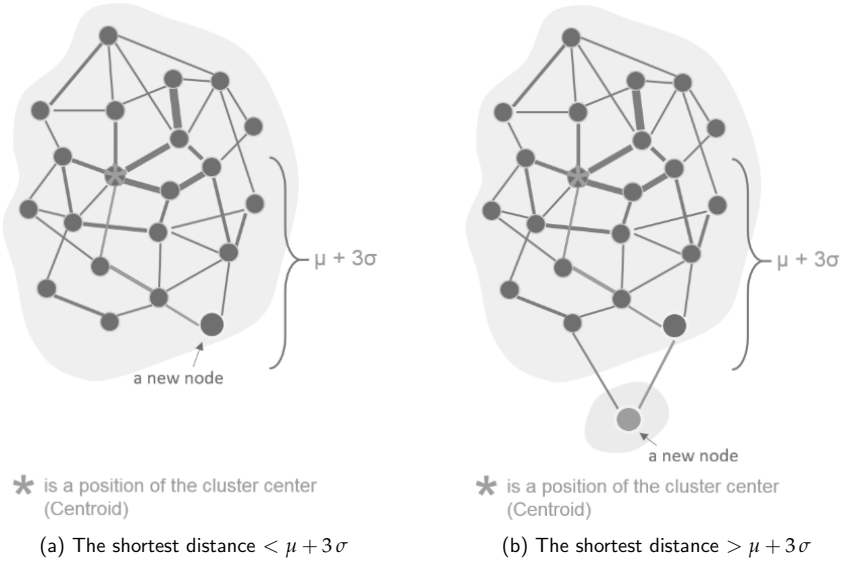
With an arrival of each node in the co-occurrence graph, a graph internal structuring process which based on a clustering building mechanism may start. This will be described below.

2.3 Cluster Building Mechanism

While documents are read, new words and relations between words are added into the co-occurrence graph, every new word find the cluster where it shall be assigned into by considering the shortest distance from the new node to the cluster center. If the shortest distance is less than the range value ($\mu + 3\,\sigma$), this word can add to this cluster, see Fig. 5a. Otherwise, this word will be added to another cluster. In case no cluster is suitable, then a new cluster will be created, see Fig. 5b.

To employ such clustering algorithm, distances between nodes need to be defined. Any two words are considered close if they often appear together, i. e. as co-occurrent. Therefore a distance d between any connected two nodes $(w_i, w_j) \in E(t)$ can be defined by

$$d(w_i, w_j) = \frac{1}{g(w_i, w_j)} \quad .$$



(a) The shortest distance $< \mu + 3\sigma$ (b) The shortest distance $> \mu + 3\sigma$

Fig. 5: Cluster building growth

To handle information in the clustering process, a local set $C(w)$ is introduced on every node with the following content:

1. $C(w) = \{w\}$,
   if the node w is created from the co-occurrence graph, building algorithm and/or the node is not a member of any cluster. From experience, to every node, an initial average distance $\mu(w)$ as well as a reasonable standard deviation $\sigma(w)$ is set.

2. $C(w) = \{w_j\}$,
   if the node w is a member of a cluster with the cluster center (centroid) w_j.

3. $C(w) = \{w, w_1, w_2, \ldots, w_k\}$,
   if the node w is a member of a cluster and is at the same time as the cluster center. In this case, $w_1, w_2, \ldots, w_k$ are the k other nodes recently assigned to be member of this cluster.

While reading documents, new nodes and edges are inserted into the co-occurrence graph $G(t)$, i. e. the knowledge of the user is changed. This may

influence the structure of this knowledge, represented by the cluster assignment in $G(t)$. The following cluster re-structuring algorithm takes those changes into account and shall be started either periodically on every node, or whenever changes in its environment occur.

1. For new or single nodes, which shall be assigned to a cluster:
   If $C(w) = \{w\}$ and there are k edges (w, w_r) with $r = 1 \dots k$ and $w_r \in W$ and $\exists\, w_i$ such that

   a) $(C(w_r) \in C(w_i))$, i. e. w is a direct neighbour of the cluster, in which it shall be inserted.

   b) $d(w, w_i) \leq \mu(w_i) + 3 \cdot \sigma(w_i)$. If there is more than one node w_i with the required properties, chose those nodes with the minimal distance $d(w, w_i)$ to w.

   then set $C(w_i) := C(w_i) \cup w$ and $C(w) = \{w_i\}$ and start the algorithm on w_i.

2. For nodes, which shall be re-assigned to another cluster:
   If $C(w) = \{w_i\}$, with $w_i \neq w$ and if there are k edges (w, w_r) with $r = 1 \dots k$ and $w_r \in W$ and $\exists\, w_i$ such that

   a) $(C(w_r) \in C(w_i))$, i. e. w is a direct neighbour of the cluster, in which it might be migrated.

   b) $d(w, w_i) \leq \mu(w_i) + 3 \cdot \sigma(w_i)$.

   then set:

   a) $C(w_i) := C(w_i) \setminus w$, i. e. remove w from the old cluster w_i;

   b) $C(w_i) := C(w_i) \cup w$, i. e. add w to the new cluster w_j;

   c) $C(w) = \{w_i\}$, i. e. set the assignment on w.

   Also, start the algorithm on w_i and w_j.

3. For nodes being a cluster centroid:
   If $|C(w)| > 1$ and $w \in C(w)$

   a) Calculate the new cluster center (Centroid) $w_n = \xi(C(w))$ in the subgraph of G induced by the nodes $C(w)$.

   b) Set $C(w_n) = C(w)$ and $C(w) = \{w_n\}$. Calculate $\mu(w_n)$ and $\sigma(w_n)$, accordingly.

c) Define X as the set of all words with
$$X = \{x | (x \in C(w_n)) \wedge (d(x, w_n) > \mu(w_n) + 3 \cdot \sigma(w_n))\}.$$

d) Check cluster quality:
Calculate the distances $d_{i,j}(w_i, w_j)$ for all pairs $(w_i, w_j) \in C(w_n)^2$.
For a given percentage p of those pairs and a constant δ:
$$d_{i,j}(w_i, w_j) < \delta \cdot \mu(w_n),$$
set $X = C(w_n)$.

e) Set $C(w_n) := C(w_n) \setminus X$ and $C(x) = \{x\}, \forall\, x \in X$.
Start this algorithm for all nodes $x \in X$. Also, start the algorithm on w_n again, if $X \neq \emptyset$.

4. STOP

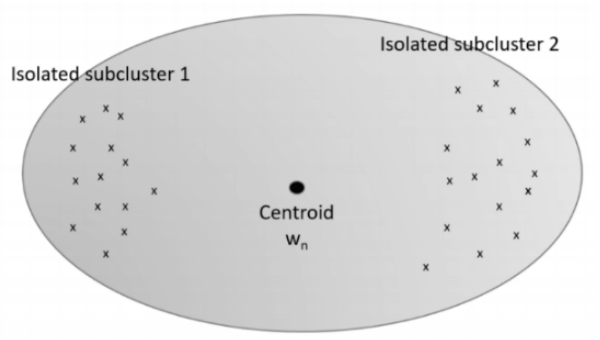


Fig. 6: Subcluster building

From step 3. d) of the above algorithm a subcluster building shall be avoided (for an example see Fig. 6). Therefore, clusters with an indicated subcluster building will be terminated and the nodes are forced to join others or build new clusters.

2.4 Hierarchical Clustering Mechanism

While documents are read, new words and relations between words are added into the co-occurrence graph, cluster building algorithm can be actioned after insertions of new words. When clusters in the co-occurrence graphs is built

completely (level 0), the hierarchical clustering algorithm will start to create the hierarchy by only reading the cluster center (node by node) of each cluster. After that, adding the first node to a new inter-cluster in the next level, then finds the inter-cluster to add every next node by considering the shortest distance from the node to the inter-cluster center. Only nodes within the distance range value ($\mu + 3\,\sigma$) from the inter-cluster center shall be a member of the respective inter-cluster. The process is repeated until only one node (root) remains, see in Fig. 7.

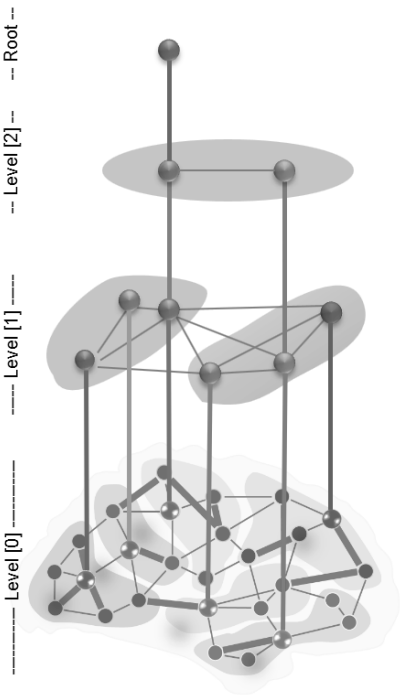


Fig. 7: Hierarchical clustering in a co-occurence graph

The hierarchical clustering in the co-occurrence graphs consists of the following steps,

1. Start with the initialization:
   a) set $l = 1$, i. e. the hierarchy start with level 1

b) set $C(0)_l = \emptyset$

c) set $w_1 := \xi_{C_1}$ and add w_1 to the first inter-cluster, i. e. start with the node which is the first cluster center (centroid).

d) set $m = 2$, i. e. continue with the next node which is the cluster center (centroid).

2. Read w_m
3. For the nodes, which shall be assigned to an inter-cluster:
   If $d(w_m, w_i) \leq \mu(w_i)_l + 3 \cdot \sigma(w_i)_l$. If there is more than one node w_i with the required properties, chose those with the minimal distance $d(w_m, w_i)$ to w_m. Then set $C(w_i)_l := C(w_i)_l \cup w_m$.
4. For nodes, which shall be re-assigned to another inter-cluster in which it might be merged.
   If $d(w_m, w_i) \leq \mu(w_i)_l + 3 \cdot \sigma(w_i)_l$ then set:
   a) $C(w_i)_l := C(w_i)_l \setminus w$, i. e. remove w from the old inter-cluster w_i;
   b) $C(w_i)_l := C(w_i)_l \cup w$, i. e. add w to the new inter-cluster w_j;
5. Update ξ_{C_i}, i. e. if any inter-cluster changing member
6. Increase $m := m + 1$
7. If $m < n$ goto 2, when n is the last cluster.
8. If $w_m \neq$ *root node* then increase $l := l + 1$ and goto 1. b)
9. STOP

As all steps have been implemented on a personal computer (PC), the results will be described in the next sections.

3 Experimentals

3.1 Goals

The presented algorithms will be evaluated in the experiments for clustering in the co-occurrence graphs. By building the clusters, the goals of the algorithm are:

1. Words and relations between words are inserted into a co-occurrence graph.

2. New words must be assigned with a cluster.
3. Cluster building algorithm can react after an insertion of new words and the evaluation of the clustering in the co-occurrence graph is updated after the words are added or moved.
4. A new cluster will be opened if there is no cluster for the new words.
5. The hierarchical clustering algorithm can be built until the last node (root) remains.

To prove this, as above indicated goals, the experiments will be described below.

3.2 Experimental Setup

The experiments have been performed on a personal computer (PC) with an Intel Core i3 CPU 550 3.20 GHz x 4 and 4 GB of RAM to show that the goals of the algorithms have been achieved. The clustering starts with only one word and one cluster at the beginning, then each new word of each sentence of each document is added into a cluster in a co-occurrence graph. The system flow processes are shown in Fig. 8.

In Fig. 8, after listing of documents are added sequentially, each sentence of each document is processed as follows:

1. *Insertion of words in co-occurrence graphs*: after stemming and stopwords removal in each sentence is processed successfully, a word, forming a linear, is inserted into a co-occurrence graph if a new word or updated data of that word is an existing word. Then, to add a relation between words which is a new connection or update the relation if this is an old connection.
2. *Clustering of words in co-occurrence graphs*: a word of each sentence must be assigned with a cluster. If a word is the first word of any cluster, then set this word to a cluster center and set the distance range $(\mu + 3\sigma)$. The distance to the cluster center ensures that each word is a member of that cluster. If distance is less than or equal to $(\mu + 3\sigma)$, then insert the word into an existing cluster. If distance is more than to $(\mu + 3\sigma)$ then a new cluster is opened for this word.
    - In case of adding new words: to check the position of the cluster center of that cluster.

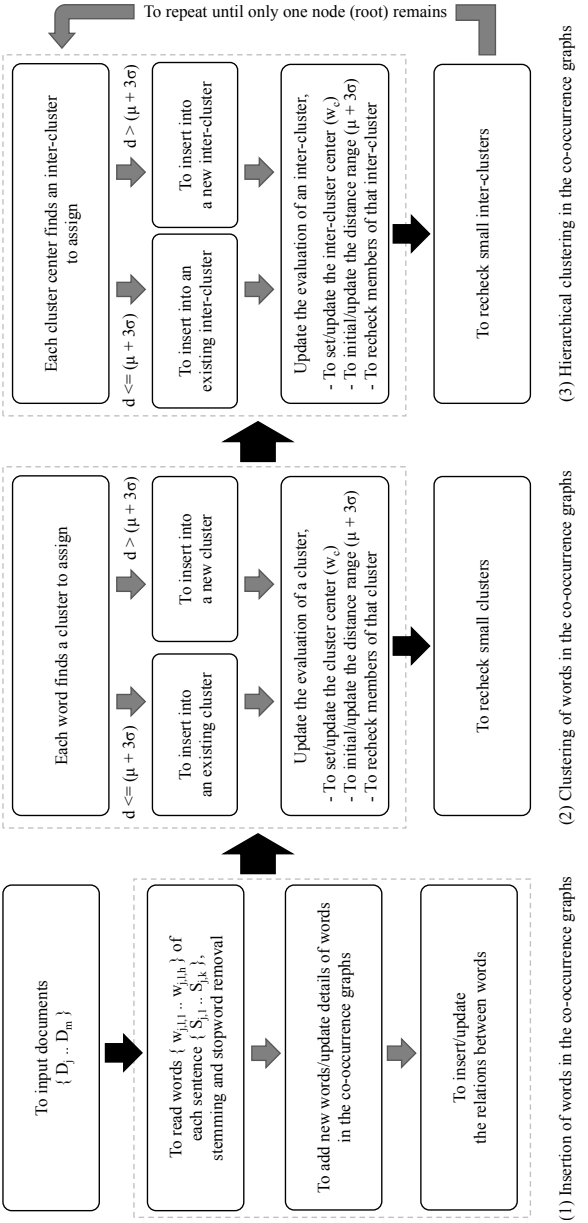


Fig. 8: The system flow processes

- In case of adding an existing word: to check previous words of this sentence that can be moved to the previous cluster that this word contains. Then, check the position of the center of the clusters that have changed the number of members.

For both cases if the cluster center change, then recalculate $(\mu + 3\,\sigma)$ and recheck the distance of members to a new cluster center. Some members will be moved to another cluster. Moreover, at the end of the insertion of each document, recheck small clusters to determine if some member can be moved to another cluster.

When all documents have been read, the clusters in a co-occurrence graph is stable then *the Hierarchical clustering in co-occurrence graphs* is processed: a word which is the cluster center of each cluster find an inter-cluster to assign. If a word is the first word of any inter-cluster, then set this word to an inter-cluster center and set the distance range $(\mu + 3\,\sigma)$. The distance to the inter-cluster center ensures that each word is a member of that inter-cluster. If the distance is less than or equal to $(\mu + 3\,\sigma)$ then this can be inserted into an existing inter-cluster. If the distance is more than to $(\mu + 3\,\sigma)$ then a new inter-cluster is created for this word. After an insertion of the last cluster center, then recheck small inter-clusters to ascertain if some member can be moved to another inter-cluster. Furthermore, to repeat all steps until only one node (root) remains.

3.3 Results and Discussion

Results of the experiment discussed herein revealed that the algorithms which have been applied to manage the insertion of words, the clustering of words, and the hierarchical clustering in the co-occurrence graphs were activated. To ensure that the algorithms of the co-occurrence graph and clustering building perform well as the cluster tends to be the same, the ten datasets (D1–D10) were set and used to construct the clustering in co-occurrence graphs. Each dataset consists of 100 documents (20 documents per topic) which covers the topic of art, car, computer, leisure, and sport as classified by online magazines articles. The experiment results were presented in Fig. 9 and Fig. 10.

In the insertion of ten datasets (D1–D10), the documents in each dataset are different but are in the same topic (5 topics) and start with one node and one cluster. Fig. 9 presents the insertion of one dataset, the number of clusters that has one, two, four, six, eight, ten and more than ten nodes after every insertion of ten documents. Overall, the small clusters will slowly disappear in the future, and the number of big clusters increases significantly.

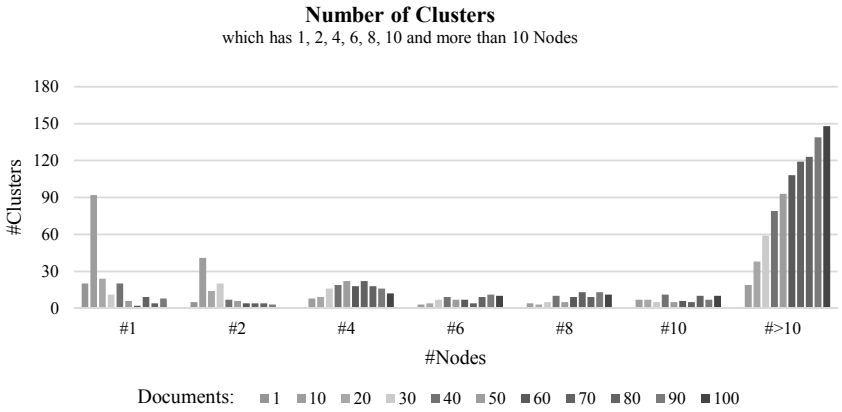


Fig. 9: Number of clusters with insertion of one dataset

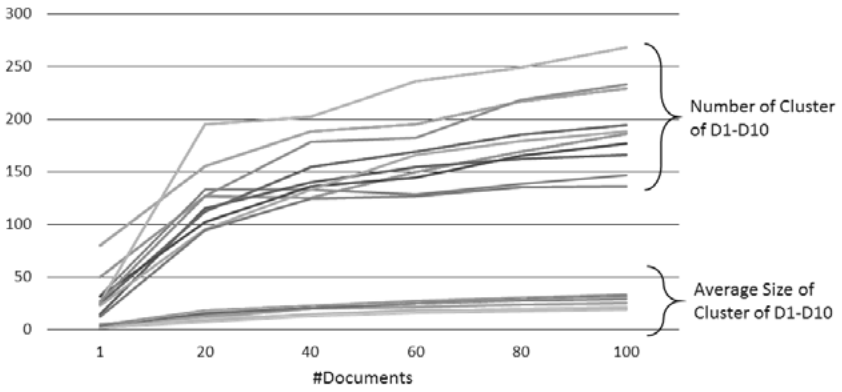


Fig. 10: Number and average size of clusters with insertion of ten datasets (D1–D10)

The graph in Fig. 10 represents an increase in the number of clusters (on the top) and the average size of clusters (on the below) with insertion of ten datasets (D1–D10). The range of values on the horizontal scale is 1, 20, 40, 60, 80 to 100 documents were read completely and on the vertical scale represents the number of clusters and the average size of the cluster. Overall, the cluster tends to be exponential.

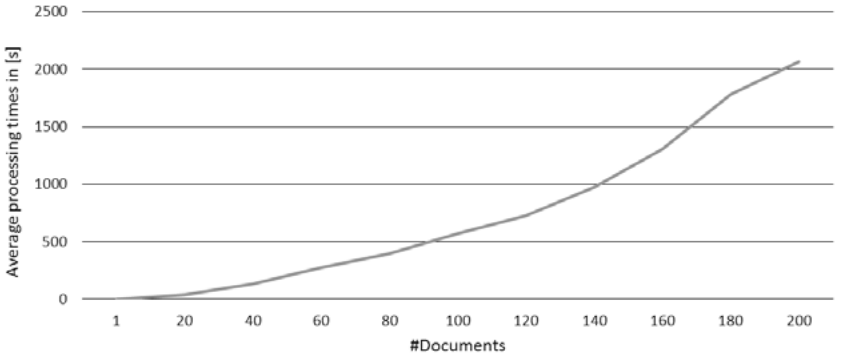


Fig. 11: Average processing times of clustering with insertion of 200 documents

As the number and size of the cluster increase steadily after starting an insertion of the first document, Fig. 11 shows the average processing time of the clusters from the insertion of 200 documents spent on each cluster which tends to increase gradually as well.

Fig. 12 presents an example of the text clustering results that represent members of each cluster. Each cluster will have one member which is designated as the cluster center i. e. there is a cluster with 28 members and "antarctica" is the cluster center. These results are called level 0 of the data hierarchy creation.

On level 0 of the data hierarchy, when a stable state is reached, the hierarchical clustering algorithm will start to create the hierarchy by reading only the cluster center (node by node) of each cluster. For this example, the cluster to be chosen for the next hierarchy level must have more than two members. Each cluster center finds the inter-cluster to assign. This process is repeated to build the next hierarchy level until the last node (root) remains, see in Fig. 13.

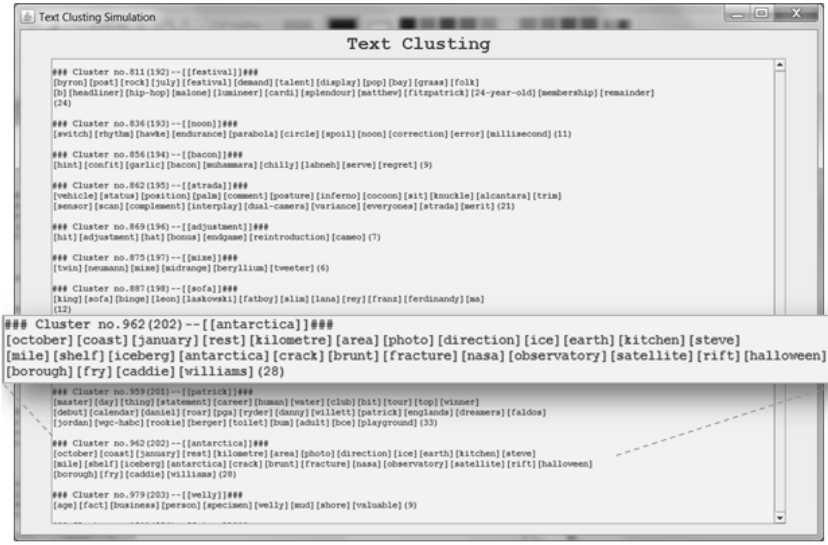


Fig. 12: Clusters building example (Level 0)

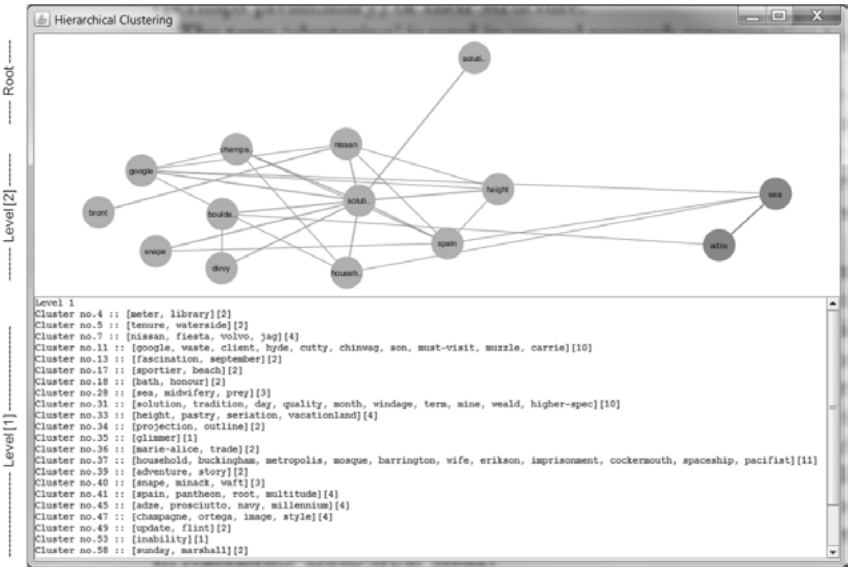


Fig. 13: Hierarchical clustering example (Level 1 – Root)

4 Conclusion

The algorithms and implementation details of the Dynamic Clustering for Segregation of Co-occurrence graphs have been presented. These development algorithms worked on one document after the other was read sentence by sentence. New words of each sentence and relationships between words were added, a co-occurrence graph was built. Each word will be assigned the cluster depending on the distance to the cluster center (centroid). Some new words were added to existing clusters whereas a new cluster was opened for adding some new words due to the absence of cluster to assign. After that each centroid was extracted and find the inter-cluster to assign. The hierarchical clustering was created and build the next hierarchy level until only one node (root) remains. Overall, this technique may be applied to (decentralised) search engines in the future.

References

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: a review", ACM Computing Survey, pp. 264–323, 1999.

[2] Campi, A. and Ronchi, S., "The Role of Clustering in Search Computing", in: *20th International Workshop on Databases and Expert Systems Application*, Linz, Austria, 2009, pp. 432–436.

[3] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections", in: *Fifteenth Annual International ACM SIGIR Conference*, June 1992, pp. 318–329.

[4] M. M. Fayal, and M. Rafi, "Clustering textual documents by extracting sequence from word-of-graph", Journal of Independent Studies and Research, Computing Volume 12, Issue 1, 2014.

[5] J. Lin, X. Li, and Y. Jiao, "Text Categorization Research Based on Cluster Idea", The Second International Workshop on Education Technology and Computer Science, IEEE, 2010.

[6] Z. Faguo, Z. Fan, Y. Bingru, and Y. Xingang, "Research on Short Text Classification Algorithm Based on Statistics and Rules", The Third International Symposium on Electronic Commerce and Security, IEEE, 2010.

[7] G. L. Baum, R. Ciric, D. R. Roalf, and others: Modular Segregation of Structural Brain Networks Supports the Development of Executive Function in Youth. In: *Current Biology*, pp. 1561–1572, `http://dx.doi.org/10.1016/j.cub.2017.04.051` (2017).

[8] L. R. Dice. Measures of the amount of ecologic association between species. Ecology, 26(3):297–302, July 1945.

[9] Normal Distributions (Bell Curve), `https://statisticsbyjim.com/basics/normal-distribution/`, last accessed 2019/05/03.

[10] M. M. Kubek and H. Unger. Centroid Terms as Text Representatives. In: *Proceedings of the 2016 ACM Symposium on Document Engineering*, DocEng 16, pp. 99–102, New York, NY, USA, ACM, (2016).

[11] Unger, H.; Böhme, T.; Kubek, M.: On Evolving Text Centroids. Published at the International Conference on Application Systems, APPIS, Gran Canaria, (2018).

Sequential Clustering using Centroid Terms

Mark Hloch[1] and Mario Kubek[2]

[1]Faculty of Electrical Engineering and Computer Science
University of Applied Sciences, Krefeld, Germany

[2]Chair of Communication Networks, FernUniversität in Hagen, Germany

Abstract: Most common text-clustering algorithms require a given a-priori number of clusters to be specified as well as the beforehand presence of all documents, which is a shortcoming for practical applications. With the work of Komkhao et. al. an algorithm to sequentially group items into clusters of unspecified numbers was introduced. This approach is now concretized by two algorithm variants using the concept of centroid terms for sequential text-clustering. First experiments show it's results and possible enhancements will be discussed.

1 Introduction

Clustering documents has always been a part of research over the last decades [1]. Focus on clustering is to group the most similar objects to appear in the same cluster. In the past, a large quantity of clustering algorithms has devised [2]. Many of the classical approaches such as the kmeans [3], kmeans++ [4] or k-NN algorithm [5] require the value k as the suggested number of expected clusters a priori. From the user's point of view this approach doesn't seem to be adequate as it requires an estimation process. Even those clustering algorithms that don't require any a priori cluster number, expose another weakness: they typically require a full set of documents. This could be addressed by using a sequential algorithm.

Apart from the classical, often used vector-based algorithms, graph-based algorithms that reflect the documents structure in a more natural, somehow brain-inspired associative way seem, to be applicable as well. The Chinese Whispers (CW) clustering algorithm which was introduced by Biemann [6] in 2006 has shown that graph-based approaches on text-clustering can be a good alternative to the classical ones. The great advantage of this algorithm lies in its linear performance, effectiveness in small-world graphs and simplicity. Additionally unlike kmeans [3] and kmeans++ [4], it does not require an initial value of k to

function properly. But the CW also comes with some disadvantages. Firstly, it tends to produce a large number of clusters with only very few terms. In addition CW, has a random property which means that for each process on the same data it will produce a different number of clusters and therefore also a varying purity making it hard to fixate it to a specific value.

With the previous work of Komkhao et. al. [9][10] a generic approach on sequentially grouping items into clusters of unspecified number was proposed. This approach is now concretized by two algorithm variants, mainly differing in their initialisation, using centroid terms as a central feature for sequential text-clustering. The primary aim of the following considerations is to get a better understanding of the behaviour of sequential clustering using centroids. First experiments show that it is possible to cluster documents in a meaningful manner without using a predefined value k provided by the user.

2 Fundamentals

2.1 Centroids

Co-occurrences can be represented as an undirected weighted graph $G(W, E)$ where each node $w \in W$ represents a term that occurs in one or more documents. A co-occurrence of two terms $w_a \in W$ and $w_b \in W$ is denoted by an edge (w_a, w_b). The edges are weighted by a significance value sig, i. e. DICE-coefficient [8], that represents the strength of the relationship of the connected nodes.

Using the concept of graphs, the distance $d(w_a, w_b)$ between two terms w_a and w_b can be defined as the reciprocal of the significance value:

$$d(w_a, w_b) = \frac{1}{sig(w_a, w_b)} \tag{1}$$

If any path (assuming the co-occurrence graph is connected) between two terms w_a and w_b is existing, the shortest path can be determined by

$$d(w_a, w_b) = \Sigma_{i=1}^{k} d(w_i, w_{i+1}) \tag{2}$$

or otherwise $d(w_a, w_b) = \infty$. Therefore, it is possible to determine the average distance $d(D, t)$ for each term t in a document D by

$$d(D, t) = \frac{\Sigma_{i=1}^{k} d(w_i, w_t)}{N} \tag{3}$$

Based on the previous considerations, the centroid term $\chi(D)$ is defined as the term with the minimal average distance to all other terms in a document D [13]:

$$d(D, \chi(D)) = MIN.$$

As shown in the publications of Kubek and Unger [11–13], the centroid term $\chi(D)$ can be used to determine the semantic distance of documents and therefore can be applied in text-clustering algorithms.

In addition to the centroid term $\chi(D)$ itself, which can be considered as centroid of first order, it is also possible to determine multiple centroid candidates in a fast way using spreading activation as shown in [14]. This is done by determining the term with the i-th shortest average distance to all other terms in a document D. As the experiments in [15] show, these candidates sometimes might be more specific than the centroid term itself. In those cases, it might be appropriate to use these candidates to find a more accurate semantic distance.

2.2 A Heuristic Clustering Algorithm

The heuristic clustering algorithm as defined in [9, 10] proposes a clustering approach without the need of an a-priori definition of the number of clusters. The algorithm itself works in a sequential manner which means that the feature vectors **F**, i.e. documents or any other data points, are processed incrementally, whereas the features then can be available during their initialisation or appear over time. The core idea of the algorithm is to compare each feature vector f against an existing set of clusters and consider its membership using a certain threshold for finding the best possible match. If there is a best match (winning cluster) the new feature is added, otherwise a new cluster with one element will be formed instead and added to the cluster model **M**.

1. **Input:** A set of input vectors f_i, $i \geq 1$
2. **Output:** A set of clusters containing the feature vectors f_i
3. **Algorithm:**
   a) **Initialisation**: Given an input vector f_1 which could have been randomly chosen $f_1 \in \mathbf{F}$ with $|\mathbf{F}| > 1$; let the cluster $\{f_1\}$ form the model M initially.
   b) **Loop**: Process all newly arriving or other existing feature vectors $f \in \mathbf{F}$ as follows:

- Calculate the membership of f in all clusters of $\mathbf{M}$
- Determine the winning cluster as the one for which f assumes the highest membership value.
- **If** the value of f's membership in the winning cluster does not exceed a given threshold,
  then associate f with the winning cluster by merging
  Â**else** extend the model by a new cluster containing just $f(\mathbf{M} := \mathbf{M}\{f\})$

The determination of a feature's membership in an existing or to form a new cluster is the most crucial part of the proposed algorithm. In general, it is suggested to use a distance measure in order to determine a membership to a certain cluster. The measure itself could be based on the distance of certain terms i.e. centroid terms [13].

In addition, a membership threshold needs to be defined, that can be either defined statically or dynamically for each cluster by using i. e. the features relation to its nearest neighbours. The latter approach is referred to the single-linkage method and is known to be susceptible to form chains of wrongly clustered features. Therefore, it is recommend to mitigate this behaviour using techniques like outlier detection, i.e. splines or other statistical methods. By detecting these outliers, it is possible to sharpen the boundaries of cluster membership and therefore a better clustering result should be expected. As the sequential clustering is an incremental process, the outlier detection shall be performed as a post-process. As the current work focuses on a first realisation, this feature is postponed as a further enhancement in later works.

3 A Sequential Clustering Algorithm using Centroid Terms

The mentioned algorithm [10] describes a heuristic approach on sequential clustering documents and provides a first outlook in how to use the algorithm in conjunction with the concept of centroid terms, in order to determine the topical relatedness of documents over a co-occurrence graph G [11]. Applied to the heuristic clustering algorithm shown in 2.2, the input vector f reduces to the dimension $|f| = 1$ by using the centroid term of each document. The sequential clustering algorithm using centroid terms and a single document as initial cluster is outlined as follows:

Algorithm a

1. **Input:** A set of documents D_i, $i \geq 1$
2. **Output:** A set of document clusters C_n, $n \geq 1$
3. **Algorithm:**
   a) **Initialisation**: Select an arbitrary document $D_{i=1}$ of existing documents D_i, $i \geq 1$, determine its centroid $\zeta(D_{i=1})$ and let it form the initial cluster $C_{n=1} = \{D_{i=1}\}$.
   b) **Loop**: Determine for each additional or newly arriving document D_i its centroid $\zeta(D_i)$ in G and:
      (a) Calculate the membership of the current document D_i for each existing cluster $C_{n\geq 1}$.
      (b) Determine the winning cluster as the one for which D_i assumes the highest membership value.
      (c) **If** the value of D_i's membership in the winning cluster does not exceed a given threshold,
      then associate D_i with the winning cluster $C_{win} \in C_n$
      Â**else** extend the model by a new cluster $C_{new} = \{D_i\}$

There are mainly three crucial parts that influence the behaviour of the algorithm which need a further investigation:

- initialisation,
- cluster membership value and
- threshold of membership (winning cluster determination).

Initialisation of the Algorithm

The initialisation of the clusters influences the quality of clustering itself by predefining a set of one or more clusters as a reference to all following documents to be clustered. The simplest initialisation is to start with a single random (or the first arriving) document as the first cluster. This approach can result in obvious inaccuracy depending on the number of documents in total to be clustered and the method of membership determination, as there is initially just one single cluster to refer to. Especially when using certain average values for membership determination in each existing cluster, it is expected that at the beginning of

the clustering process a decision regarding cluster membership is inaccurate for small cluster sizes. This variant of initialisation is denoted as *algorithm a*.

An alternative option for initialisation is the concept of antipodean documents [10]. This approach results in at least two clusters of the most distinct documents in a set of existing documents. These two clusters of antipodean documents are most likely to be topically unrelated. A basic requirement in general is, that there exist at least two documents during initialisation of the algorithm. Additionally, it must be noted that further arriving documents (as they may occur in an online clustering algorithm) are not considered for being used for initialisation, even if they might be more antipodean than the used ones. The algorithm variant using antipodean documents for initialisation is denoted as *algorithm b* and outlines as follows:

Algorithm b

1. **Input:** A set of documents $D_i,\ i \geq 2$
2. **Output:** A set of document clusters $C_n,\ n \geq 2$
3. **Algorithm:**
   a) **For each** existing document D_i ($i \geq 2$) determine its centroid term $\zeta(D_i)$ using spreading activation [14]
   b) **Loop:** Determine the shortest path between the current document D_i and all of the other documents; recognize the most distant two documents D_a and D_b
   c) **Initalisation:** Let the D_a and D_b form the two initial clusters $C_{n=1} = \{D_a\}$ and $C_{n=2} = \{D_b\}$.
   d) **Loop**: Determine for each additional or newly arriving document D_i its centroid $\zeta(D_i)$ in G and:
      i. Calculate the membership of the current document D_i for each existing cluster $C_{n\geq 1}$.
      ii. Determine the winning cluster as the one for which D_i assumes the highest membership value.

iii. **If** the value of D_i's membership in the winning cluster does not exceed a given threshold,
then associate D_i with the winning cluster $C_{win} \in C_n$
Â**else** extend the model by a new cluster $C_{new} = \{D_i\}$

As the time-complexity of the Dijkstra's algorithm depends on the number of edges E and nodes N of G, it might be useful to reduce the number of calculations in step 3b by limiting the number of considered documents. I. e. one could choose an appropriately large, but not too large, subset of random documents. In addition, some documents may have the same centroid term where computation time can be saved with a caching mechanism.

Cluster Membership Value

The sequential clustering algorithm using centroid terms requires for each new document to be matched against the existing clusters C_m. Therefore, the average distance between a document centroid and all centroid terms of a cluster is determined as suggested in [10]. The distance between the centroid and each of the centroids in the clusters is determined using Dijkstra's shortest path algorithm. First experiments show that depending on the number of documents there can be a performance issue having documents with the same centroid term. This results in unnecessary calculations and thus can be optimized using a caching mechanism. For this purpose, a look-up table is built up and used during the membership determination:

1. **Input:** Create an empty look-up table *LT*; Key value is the hash of two concatenated terms; value shall be the distance (shortest path) between these terms
2. **Output:** The distance (shortest path) of a documents centroid and a clusters centroid d_{D_i,C_n}
3. **Algorithm:**
   a) **For each** unclassified document represented by its centroid term C_{D_i}

      For each centroid term C_n in the current cluster

      (a) Form the concatenatenation $CC_{i,n} = D_i C_n$

      (b) Determine a hash value $hash = md5(CC_{i,n})$

      (c) Lookup *hash* in *LT*

(d) **if** $LT(hash)$ exists, use its distance value for further processing **else** determine the distance d_{D_i,C_n} using Dijkstra and create a new lookup entry extending LT

Threshold Determination

Good results in terms of clustering documents of similar topics require a mechanism that is capable of judgeing about the membership of a document to a cluster. Thus, the role of the threshold directly influences the number of clusters and the cluster results. If the threshold is set to high, the number of erroneously assigned documents to a cluster will increase while the number of clusters will be reduced. If the threshold is choosen too low, it would lead to a higher number of clusters - in the worst case - for each document a single cluster. In general, there are two approaches using a threshold: static or dynamic.

A static threshold can be easily set by the user based on general experience and by using a graphical interface or may be predefined in any other way, programmatically. Using a graphical interface requires the user to interact and is therefore of more interest in conjunction with searching. An unsupervised suggestion of the cluster for a given set of documents seems more useful, i.e. for topic detection or recommender systems. In general, a static threshold does not seem to be easily found as the corpus itself varies in its individual properties. If there is a static value, it might by chance fit for a corpus but for another it would lead to a worse clustering result.

The better approach is the dynamic adjustment of the threshold according to the local connections of the centroid with it's nearest neighbours which is also suggested in [10]. As the centroid needs to be known before clustering this process can be easily expanded by performing a breadth first search starting at the individual centroid in order to find its nearest neighbours. For larger document sets the performance of the algorithm can be greatly influenced by nearest neighbour determination.

Instead of identifying any available nearest neighbour of a centroid term, a breadth search is performed and such nearest neighbours that don't exceed a certain allowed maximum path length are taken into account. Based on several tests, a maximum path length of 20.0 has been found to be appropriate in terms of performance and accuracy in corpora containing up to 10.000 nodes. Further calculations in order to find a rule for a more precise determination of the path length or the derivation of a range of path length values depending on the number of nodes and edges will be subject of future research.

4 First Experiments

For all of the following experiments, natural language preprocessing in form of sentence extraction, stop word removal and baseform reduction has been applied to all of the documents. Then the co-occurrence graph G has been build on sentence based co-occurrences. The distance between the terms is represented by the reciprocal value of the significance value which is determined using the DICE-coefficient [8].

In order to obtain a consistent single graph G, all small sub-graphs from the original co-occurrence graph have been removed by determination of the largest connected sub-graph within G. Technically the graph is stored in a embedded graph database using Neo4j[1] where the nodes represent each term and the edges are annotated with their significance and respective distance value.

4.1 Additional Post-processing using Chinese Whispers Algorithm

The Chinese Whispers algorithm [6] is due to its performance and good clustering results used as the state of the art reference for the conducted algorithms. The given implementation of the Chinese Whispers algorithm and its behaviour itself required some additional post-processing of the results which is subject to the next paragraphs.

Category Mapping

The available implementation of the Chinese Whispers algorithm produces clusters of terms but provides no information which documents are related to these terms. In contrast the implementation of the sequential clustering algorithm produces clusters of documents. In order to make the algorithms comparable, the resulting term-clusters of the Chinese Whispers were transformed into document-clusters as follows:

1. **Input:** A set of documents D_i, $i \geq 1$
2. **Output:** A set of document clusters DL_{C_n}
3. **Algorithm:**
   a) Perform CW Algorithm (50 iterations);
   Result: n clusters C_n containing the terms t_i

[1] `https://neo4j.com`

b) **For each:** document used for clustering determine its term vector $\mathbf{f_D}$ of t_x terms existing in G

c) **For each** cluster C_n; set $DL_{C_n} = \{\}$

For each term $t_i \in C_n$

Determine the according document name DN by comparing t_i against all $\mathbf{f_D}$.

Insert document name in $DL_{C_n} = DL_{C_n} \cup DN$

For each document name in DL_{C_n} determine its category[2]

Set the category as the winner for the term with the highest category count

Reduction of Result Clusters

In addition to the previous considerations, the Chinese Whispers algorithm is known to create many clusters in real world scenarios [6]. Without any filtering some pre-tests showed that for the used corpora it was likely to result in more than 800 clusters with very few terms.

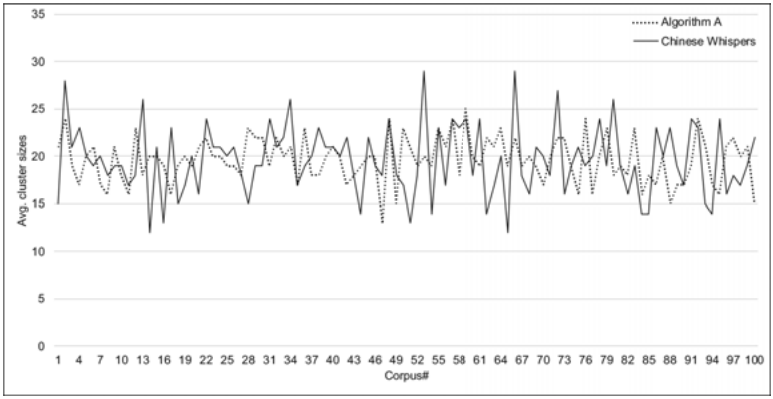


Fig. 1: Resulting average cluster sizes for algorithm a vs. Chinese Whispers

[2]Each category is part of the document name tagged by the author

In order to reduce the number of resulting clusters only those clusters were taken into account with a number of members that was at least equal to 40% of the maximum avg. cluster size of the corpora. The remaining clusters, for Chinese Whispers clusters 21 in average, were used for further observations.

Figure 1 shows exemplary the avg. the cluster sizes for all examined corpora for algorithm a vs. Chinese Whispers. The total avg. cluster size for algorithm b was, with an value of 20 clusters, identical to algorithm a.

4.2 Document Clustering

The conducted experiments were performed on 100 corpora each consisting of 100 documents. Each of the corpora consists of a random number of 25 documents of the categories *politics*, *cars*, *money* and *sports*. The documents themselves were German online articles from "Die Welt" website, each tagged by the author with their respective text category. This human tagging is further used as a gold standard in order to evaluate whether a document has been clustered correctly or not.

The experiments have been performed using the algorithm a with a random selected document as initial cluster. In addition, algorithm b using the concept of antipodean documents (see section 3) for initialisation was tested. As a reference for all corpora, the Chinese Whispers algorithm has been run. The cluster membership has been determined as described and of the respective threshold was calculated dynamically as described in section 3.

Even though the algorithms a and b could be examined during the sequentially processing of the documents, the conducted experiments focus on the entirely processed set of documents, and therefore the purity has been determined for the entire corpus. Based on counting the documents of the most common class for each cluster, the purity is calculated by

$$purity(C, M) = \frac{1}{N} \sum_{k} \max_{j} |c_k \cap m_j| \tag{4}$$

with N as the total number of documents, the set of clusters C and M as the set of classes.

Figure 2 and Table 1 show the purity [7] of algorithm b is 0.72 vs. a purity of 0.70 for algorithm a. It shows that the determination of two distinct documents as initial clusters results in a clustering of higher quality.

Table 1: Average purity for each of the tested algorithms

Algorithm A	Algorithm B	Chinese Whispers
0.70	0.72	0.75

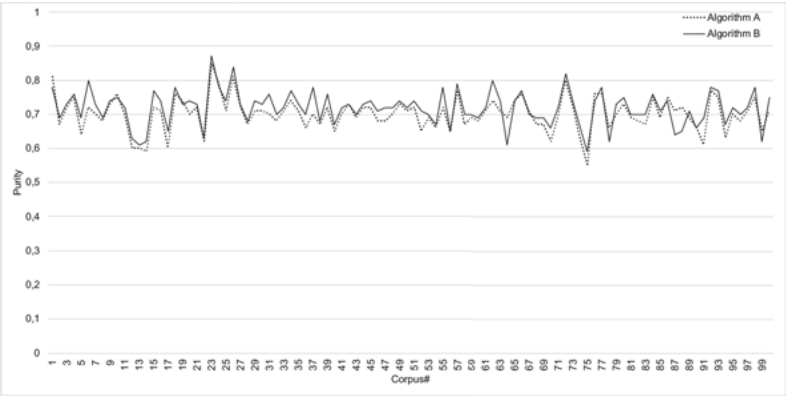


Fig. 2: Purity of algorithm a vs. algorithm b

In comparison to the Chinese Whispers algorithm, both implemented algorithm variants are not as good but already within reach. Considering that there was no outlier detection or other optimization, the results show that the approach of sequential clustering using centroid terms is sensible For the outlier detection itself [10] suggests the use of splines as an alternative option for further examination, one might also think of using the three-sigma rule.

In contrast to Chinese Whispers, the result of each run of algorithm a and b is not changing for a set of documents. The good results of Chinese Whispers therefore only provide a snapshot that might have a higher or lower purity.

Execution Time of Algorithms a and b

Figure 3 shows the execution time of clustering each of the tested corpora. It can be seen that algorithm a is much faster than algorithm b. The main reason is the time-consuming process of finding the two antipodean documents. In the conducted experiments, the distance of all 100 documents to each other was determined. A better approach for future implementations is to use a subset of these documents instead, i. e. by selecting them randomly.

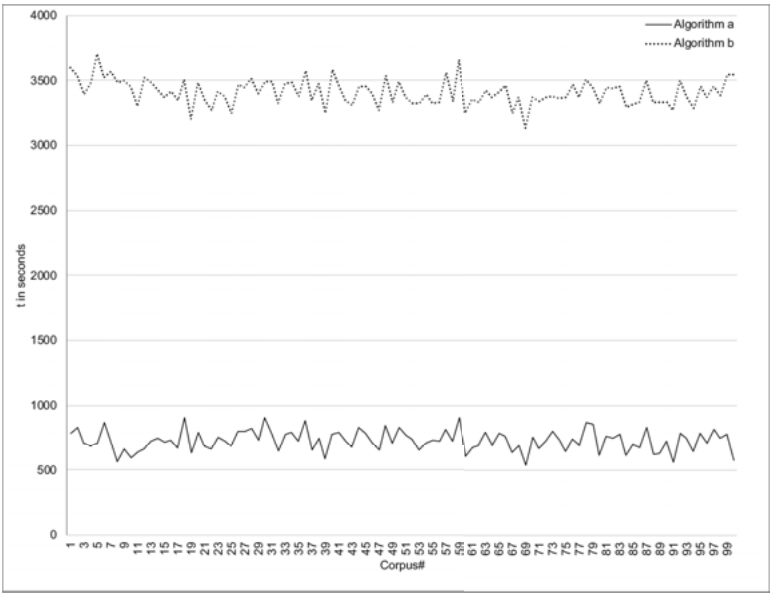


Fig. 3: Execution time algorithm a vs. algorithm b

In term of execution time, Chinese Whispers is due to its linear performance still unbeatable. The concept of antipodean document determination and centroid term detection are too time consuming for real-time clustering in interactive scenarios and therefore more suitable for offline applications.

4.3 Suggested Enhancements

Membership Determination

First experiments show that the clustering is often imprecise, especially at the start of the algorithm. The clusters themselves contain only a low number of documents and therefore, at the beginning, the membership estimation is based on rather rough average values. In time, when the clusters start to grow, the average values for each cluster become more accurate. This behaviour can be mitigated i. e. by performing a second run for the same 5–10 % of documents again (reclustering).

4.4 Clustering using Centroid Candidates

The conducted experiments using algorithm a and b show that there is still room for improvement to increase the purity towards Chinese Whispers. The above experiments so far have been using only the first order centroid terms to reflect the semantic distance of the documents and their clustering. With regard to the work of [15], it might be useful to use the centroid candidates in some cases as it shows that candidates sometimes can be more specific to represent the actual meaning of the document.

This property is actually reflected by the degree of a terms node: terms with a high degree typically have a more general meaning, terms with a lower degree more specific. It is to be expected that the first order candidates in average will have a lower degree and hence a higher specificity.

In order to confirm this fact, algorithm a has been run for the 100 corpora once using the centroid term and additionally using the second order centroid candidate. For each of the corpora the average degree of the centroids and candidates were determined.

Figure 4 shows that the degree of the first order candidate run shows constantly lower degrees compared to the second and third order centroid candidates.

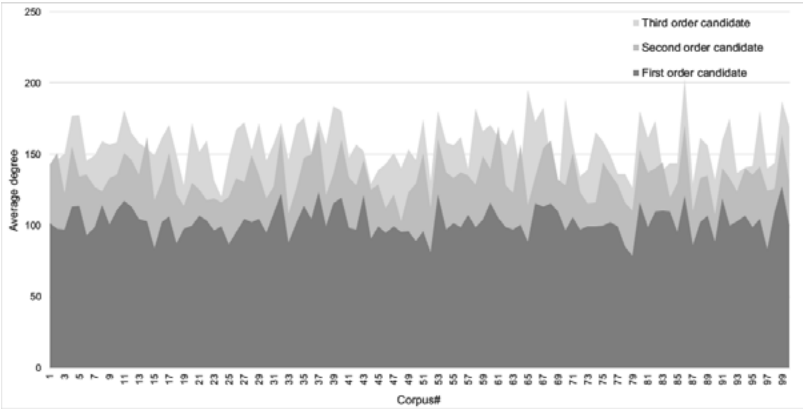


Fig. 4: Average degree of first, second and third order candidate

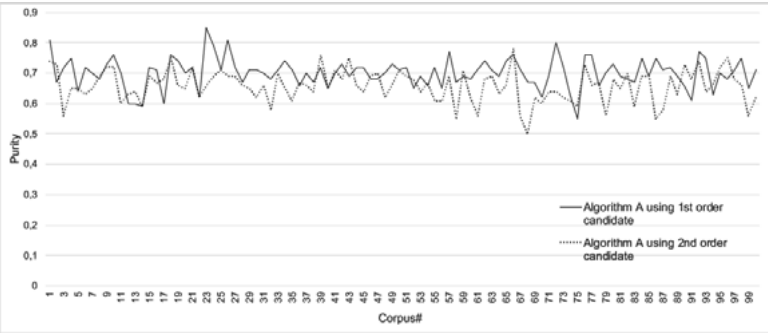


Fig. 5: Purity of algorithm A with first order centroid vs. 2nd order candidate

Hence, the centroid run uses more specific terms during the clustering process resulting in the above shown table 1. As the second order centroid candidates are more unspecific, their usage is expected to result in a lower purity than the usage the first order centroid terms for clustering. Figure 5 shows the comparison of the purity values for algorithm a using the first order centroid vs. algorithm a using the second order centroid.

It can be observed that in general the use of the first order centroid term results in better clustering for most of the corpora. But it can also be observed that for several corpora – in fact approx. 20 % – the purity by using the candidate seems to provide a higher purity (Fig. 5).

Referencing again figure 4, one can notice that the average degree of selecting the lowest degree out of the candidate and second order candidate appear to be more specific than the primary centroid candidate. As an enhancement of the clustering algorithm, this property – in context of the previously made observations – results in the following future approach which might also be applicable:

1. Determine the degree of each centroid and candidate for each document
2. **if** degree of centroid candidate $i >$ candidate $i+1$ then use candidate $i+1$ **else** use the i-th centroid candidate

This means instead of only relying on the centroid terms for clustering the algorithm should make a decision based on the degree and therefore on the specificity before each clustering is performed. Future work will be related in concrete implementations of this proposal for algorithm a and b.

5 Conclusion

Most text-clustering algorithms require a given a-priori number of clusters to be specified as well as the beforehand presence of all documents. It was shown that a graph-based realisation of sequential clustering algorithm using centroid terms is applicable for text-clustering. In order to obtain first experience two different variants of the generic algorithm were realized. Firstly, a straight-forward approch using a single document to initially form a cluster was presented. Secondly, an algorithm that performs additional pre-processing by determining antipodean documents was introduced which therefore forms two initial clusters at start time. The latter shows purity-wise and without using any further optimization good results already.

It was shown that there is also space for performance and cluster quality enhancements. By performing additional re-clustering considering the dynamic aspects of speech i. e. changing relationships between terms could lead into a improvement of cluster quality and will be suspect of further research. In terms of execution time, it seems to make sense to shorten the determination of antipodean documents by using a subset of the totally existing documents. In addition, first experiments show that the combination of different levels of centroid candidates may lead to better overall clustering results. Therefore, further investigations will focus on these optimization steps including experiments with a more fine-grained threshold and other parametrisations.

References

[1] Vladimir Estivill-Castro. Why so many clustering algorithms: A position paper. *SIGKDD Explor. Newsl.*, 4(1):65–75, June 2002.

[2] Rasmussen, E. Information Retrieval: Data Structures and Algorithms. Prentice-Hall (1992)

[3] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A K-Means clustering algorithm. *Applied Statistics*, 28(1):100–108, 1979.

[4] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[5] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[6] Chris Biemann. Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, pages 73–80, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[7] Manning, Christopher D. and Raghavan, Prabhakar and Schütze, Hinrich Introduction to Information Retrieval *Cambridge University Press*, 2008

[8] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[9] Maytiyanin Komkhao, Mario Kubek, and Wolfgang A. Halang. Sequentially grouping items into clusters of unspecified number. In *Advances in intelligent Systems and Computing*, 2017.

[10] Maytiyanin Komkhao, Mario Kubek, and Wolfgang A. Halang. Sequential clustering and condensing the meaning of texts into centroid terms. *Information Technology Journal*, 2018.

[11] M. Kubek, T. Böhme, and Herwig Unger. Empiric experiments with text representing centroids. In *6th International Conference on Software and Information Engineering (ICSIE 2017)*, 2017.

[12] M. Kubek and H. Unger. Centroid terms and their use in natural language processing. In *Autonomous Systems 2016*. VDI-Verlag Düsseldorf, 2016.

[13] M. Kubek and H. Unger. Centroid terms as text representatives. In *Proceedings of the 2016 ACM Symposium on Document Engineering*, DocEng '16, pages 99–102, New York, NY, USA, 2016. ACM.

[14] Kubek, Böhme, and Unger. Spreading activation: A fast calculation method for text centroids. 2017.

[15] M. Kubek. *Concepts and Methods for a Librarian of the Web*. FernUniversität in Hagen, 2018.

Dynamic Generation of Links and Forwarding to Related Web-based Content

Georg Philipp Roßrucker

Abstract: This paper is dedicated to artificial link generation based on methods and state of the art scientific approaches of natural language processing and information retrieval (NLIR). The motivation to conduct research in this field is presented. It is driven by the vast growth of the web and the complexity of its structure, which can no longer be managed manually and makes computer-aided analysis imperative. Additionally, the predominant market power and influence of a few global players offering popular web services to mainstream Internet users are a driving force behind the author's desire to research. In order to overcome the status quo, a solution outline and proof of work are drafted. The solution outline illustrates how NLIR methods can potentially be applied and combined with others in new ways to achieve the goals of strengthening and improving the quality of interconnectivity and liberalization of the Internet. Its final shape will crystallize in the course of subsequent research. The presented proof of work for link generation is designed as the first building block for the desired comprehensive solution and is oriented towards the internal linking of Wikipedia as a test field and benchmark. Conceptual considerations on the process of link generation unveil some of the challenges to be met: the use of NLIR methods in order to determine suitable link targets, the different types and the right placement of links, long term aspects, such as forgetting and learning new relationships over time, and many more.

1 Introduction

1.1 Background

Since the "World Wide Web" (WWW) was invented in 1989 it has been on an exceptional growth path [1, 2]. The early WWW could be described as a manageable directed graph of websites that were connected by hyperlinks. Hyperlinks still are the key technology allowing users and software to navigate

through the web by providing a model for the relationship between resources [3]. Eventually, with a growing number of websites, it became difficult, to almost impossible for webmasters to keep a complete view of all resources over time and maintain the hyperlinks to all the relevant sources. According to [4] at the time this thesis proposal was written the number of websites[1] approached *1.7 billion*.

Web catalogs were an early approach to collect and categorize websites, by storing links and meta-information in their databases. These catalogs made it easier for users to find relevant content by searching and filtering for the information needed. However, this approach was limited to the scope and size of the respective catalog so that users may have needed to consult multiple catalogs at the same time. The resulting hypertext network structures became subject to research regarding the rating and ranking of websites, e. g. in [5, 6]. This ultimately led to the development of powerful web search engines which employed algorithms to sort and rank websites with respect to specific search terms.

1.2 Motivation

The rise of keyword-based search engines has revolutionized the web and made it more convenient for users to satisfy their information needs. Information became easily accessible through well-chosen search queries and ever-improving search algorithms. As of today, the most popular web search engine is Google [7]. A hypothesis, that yet needs to be tested is, whether or not the success of Google has changed the shape of the web and actually led to fewer links and a weaker connection of the web graph. An initial review of secondary data did not result in a clear response [8–10].

The author of the proposed thesis does not intend to question the efficiency or the quality of today's leading search engines. It is dedicated to the question of how information retrieval in the WWW could be liberalized, designing an alternative approach and making it available and accessible to the general public. The motivation for doing so arises mainly from the following three considerations, assessed as most critical by the author.

- *"Single Points of Failure"* arise when single components can cause the failure of the whole system, they are part of [11]. For web-services, this can arise when they are offered by a single or only a few predominant providers. In the context of information retrieval in the ever-growing WWW,

[1] Number of websites is defined as number of unique hostnames

this means that resources can hardly be found anymore without the use of search engines. A failure of Google would, therefore, make it difficult for most and even impossible for some users to find information sources online. Looking beyond information retrieval it has been shown, that the failure of central web-services is not a misconception, but can actually have a significant impact on the web [12].

- The *concentration of power and control*, with respect to information retrieval, under the control of a few providers, should be questioned. They control the results which will finally be displayed to the users. The concerns that this power may be misused are strengthened by mostly non-transparent search algorithms, which cannot be analyzed and validated publicly. Additionally, it cannot be ruled out that providers misuse their power in order to influence people in a political or ideological manner. This may also happen unintentionally e.g. through filter bubbles [13]. Politics and legislation may also have an influence on how algorithms are implemented and how results are rendered. Legal censorship may lead to biased or shortened results in certain countries [14]. Last but not least, prediction and personalization of search results are often based on the creation of detailed user profiles. This allows service providers access to sensitive personal data which, for various reasons, should be considered critical.

- The *economic implications* of a few big players should be considered. Since today Google and others are de facto gatekeepers to the web, they are in charge of allowing businesses to appear in search results, giving them power over whether these businesses can do business or not. This makes online businesses highly dependent on these providers and puts them in a weak position. They may need to adapt business processes and values to comply, or pay money for visibility, in order to avoid putting their own business at risk. This influences business models and restrains entrepreneurial freedom of small and medium enterprises, who have no leverage against the power of the big players. Search results may also be driven by the provider's own business interests. Due to a lack of transparency, it is not clear how conflicting search results will be handled. EU antitrust rulings, e. g. against Google show the relevance of this problem [15].

1.3 Intended Results and Tasks

The solution to be derived during the course of this research should help to strengthen the interconnectivity of the web graph and reduce the dependency

on centralized information and search providers. This shall be achieved through a dynamic link generating system and a mechanism for user forwarding to relevant web resources. The findings shall serve as a proof of concept, but also help to design corresponding marketable web-services.

In order to validate and compare existing and new approaches, one of the first tasks will be the definition of measures and indicators in accordance with the objectives. These measures should include for example the efficiency of information retrieval, the degree of interconnection, time series data of interconnection and the quality of link generation.

Developing a proof of work for such a service is the main task in order to demonstrate a competitive alternative to the predominating search services. This needs to be split up in two sub-tasks. First is to conduct research of the literature and a theoretical discussion of findings. The following list of topics will be covered:

- Collecting and retrieving information sources in the WWW
- Methods of natural language processing, categorization, and clustering of resources
- Storage of data (including information sources, metadata, and dynamic links)
- Output and visualization of results and hyperlinks to information sources

The technical implementation, which gradually evolves towards the final proof of work, is the second sub-task. It will evolve from the implementation of isolated software modules towards fully self-contained approaches, which will later be enhanced towards a distributed solution.

Regardless of the final shape, the quality of the resulting solution needs to be evaluated and compared with existing and alternative approaches. This should be undertaken on the basis of the measures to be defined. A final discussion of the results shall clarify whether or not the goals of liberating the web and strengthening interconnection could be achieved. Further study may be proposed in order to clear unresolved questions.

This paper covers a comprehensive discussion and analysis regarding link generation and forwarding in the WWW. The current state of the art in surrounding areas of science and technology will be presented. The hypothesis that

interconnectivity may have declined is approached by considerations on interconnectivity and the nature of hyperlinks. Finally, a solution outline will be developed, which consists of a general solution idea and proof of work for link generation.

2 State of the Art

The proposed thesis will come in contact with several areas of scientific research. In particular, these consist of information retrieval and search engines, natural language processing, and last but not least Peer-to-Peer (P2P) networks. These three areas will become subject to detailed analysis in order to understand their evolution over time and their current state of the art. This will be the foundation for the development of new methods and models. The following paragraphs roughly summarize the information gathered through a brief review of the literature.

On the fields of information retrieval and search engines, it will be crucial to look closer into the mechanics of existing search providers. According to [16] Google is today's leading search engine, which is why its algorithms and methods should be considered as a benchmark for others. In [7] the working principles of Google's original page rank algorithm are presented. Beyond that, the working principles of other well-established search engines such as Bing or Baidoo, and more recent approaches, like the "WebEngine", will be analyzed too. The "WebEngine" is a P2P-based, distributed search engine, introduced in [17].

Besides resources of information nature, transaction-based web services play an important role in today's online ecosystem. Since many services provide transactions rather than information, one also needs to consider the users' intent of surfing the web. Research in this area was for example done in [18]. This helps to continuously improve algorithms and develop methods connecting content of all kinds while taking the users' intent into account.

On the task of creating relations between web resources, it is essential to understand the state of the art in natural language processing. Optimized algorithms enable automatic analyzation of the contents and contexts of huge amounts of text documents. Lexical analysis, n-grams, and collocations are examples for methods that allow to analyzing and comparing text corpora. Yet, their performance is weak when they are used to compare or cluster texts. This is especially the case when a different wording is used or the order of words matters. Other

approaches, such as Text Centroids and doc2vec, follow a more sophisticated approach.

"Text Centroids" build upon the physical model of the center of gravity and apply this to text corpora. They were introduced and discussed in [19, 20]. "Word2Vec", introduced in [21], allows comparing words through their vector representations in multidimensional spaces. This concept was further developed to "doc2vec" in [22], which generates "Paragraph Vectors" for combinations of words within sentences, paragraphs or whole documents. Both approaches have in common that they do not look only at the bag of words but also consider the proximity in terms of their meaning and the order in which words are used in a given text. This can make a significant difference in interpretation and comparison.

Sharing content and resources independently from organizations or companies are key drivers of the author's motivation. Therefore, a way to achieve distributed and independent generation and provision of contextual information and metadata of web-based content needs to be developed. Literature shows that a great variety of web services, related to information and data-sharing, employ P2P protocols for this purpose. Examples are financial services, video on demand services, or lookup protocols [23–25]. In [26] it is shown that locally running algorithms can optimize node connections, content distribution, and traffic flow on P2P networks. Consequently, they influence the global shape of a network. Methods from this and other studies are likely to be employed in the context of this work.

In [27] a method to dynamically build links between text documents is proposed. An extension to the HITS algorithm [5] was developed which allows comparing a document's list of "hub"-terms with another document's list of "authority"-terms. If the similarity between these lists exceeds a certain threshold a directed link from the first to the second will be suggested. This approach may serve as a performance measure for the idea introduced in this work.

3 Interconnectivity of the WWW

3.1 Evolution of Interconnectivity

The motivation of this research project is, among other things, driven by the hypothesis that interconnectivity in the WWW has declined with the rise of se-

arch engines. Trying to give a meaningful answer, the assumption should be refined and additional aspects taken into account.

Data shows that the rapid development and growth of social networks has particularly impacted the web in the 2010s. Billions of individual users were suddenly able to exchange information, links, and individual content with online communities [28]. In 2019, social network penetration is expected to reach 72 percent of the worldwide online population [29].

This suggests that the rise of the most popular networks such as Facebook, Twitter, and Instagram has, to some extent, changed the use and the shape of the WWW. With regard to hyperlinks, social networks are suitable tools to substitute link lists, personal homepages, and web catalogs. Since social networks heavily rely on user-generated content and interaction, they have created an all new source of hyperlinks. Also, they offer a much more dynamic and interactive method to create, share, and display hyperlinks.

To conclude, the sole analysis of the number of links in relation to nodes, i. e. websites, over time, does not take into account the change in the nature and source of links. Ex-ante the era of social networks, direct links between websites and web catalogs were the primary source of links. Since this has changed, the nature and source of links should be considered in a time series analysis, accordingly. A distinction should also be made between closed networks, where links are only accessible to members, and open networks, where they are freely available and accessible to everyone.

3.2 Measuring Interconnectivity

Crawling the whole web and creating reliable measures will be a difficult task. The sheer number of websites, which is approaching 1.7 billion by the time of this writing [4], and an even greater number of hyperlinks call for enormous computing power and highly efficient algorithms to generate a reliable graph. Additionally, the scope of the graph is critical for the analysis of the source and nature of links. For that reason, all open and closed parts of the network, e.g. the aforementioned social networks need to be observed.

Instead of crawling the entire web it seems more feasible to approach the problem by statistical means. If a method can be found to create an independent and identically distributed (iid) sample of the web it will be possible to determine statistically significant measures for the aspects mentioned above.

For a present analysis, a qualified sample could be rendered by applying proper crawling methods, fulfilling necessary iid criteria. Time series analysis, on the contrary, will be a more difficult task since multiple and equally arranged iid samples need to be rendered for several points in time. How these samples will be derived should be investigated in more detail.

Utilizing existing web graphs would reduce efforts of generating samples in comparison to crawling the web. However, similar iid criteria for sampling existing graphs need to be applied. For time series analysis, it should also be ensured that the graphs are reliably and consistently created over time. For historical data, existing web crawls, such as the common crawl dataset [30], could be examined. If they turn out to be suitable, i. e. were rendered by applying identical crawling methods over time, they could be considered for past and present sampling, making the results comparable.

4 Solution Outline

4.1 Solution Idea

The development of the anticipated solution is divided into several steps building upon each other. The first step will be the evaluation and identification of *entities* that are involved in the scope of dynamically building, providing, and consuming links between web resources. The second step will cover the evaluation and definition of *tasks and functions* performed by these entities. The last step will be the *implementation* of several approaches.

Entities

From an ex-ante perspective, the following entities will be part of a solution and therefore need to be evaluated. This list may be extended or subdivided over the course of research:

- *Information Sources / Content Providers*, e. g. web servers and websites
- *Individuums*, i. e. users who are represented by the software running on client hardware such as desktop computers or mobile devices
- *Intermediaries* who, besides the two aforementioned, provide all or some of the functions evaluated in the second step

Tasks and Functions

In an effort of creating dynamic links between information sources, the entities need to execute several functions.

In order to build a database of available resources, source seekers, need to reveal and collect new information sources. They could be implemented as autonomous programs running on servers or clients, or being triggered by user activities. All methods need to be evaluated with regard to their reach, relevance and also privacy concerns. Examples for these methods are:

- Querying search engines for keywords, extracted from already known resources
- Web crawling starting from already known resources until a defined depth is reached
- Analyzing inbound traffic's HTTP-referrers
- Tracking users' activities while they are surfing the web

Analyzers or "Reasoners" need to be in place and serve the purpose of analyzing the content of newly discovered and already known information sources. They will apply methods known from natural language processing in order to *find keywords*, *contexts*, and *topics* for a given information source. Information sources should also be *clustered and concatenated* in order to determine similar topics, source topics (generalizations), or child topics (specializations).

Information sources covering the same topic with a similar level of detail need to be *ranked* so that only relevant and most promising links are generated. The ranking needs to be an ongoing, repetitive task that allows refining the list of top related resources over time. The definition of quality characteristics and application in a ranking scheme need to be studied.

The solution will require *data storage* to store all relevant information about known resources: URLs, basic metadata, context and cluster information, ranking, and most importantly the dynamically generated graph of links that connect them.

Visualization, i. e. displaying connections between resources, is an essential part of the solution. One option to realize this could be to embed links into the pages' contents. This could be implemented by a server- or client-side software extension. Another option is to generate dynamic subsites or APIs, which need to be

accessed by consumers and will otherwise remain invisible. Besides static hyperlinks, it should be considered to parameterize the forwarding, e. g. allowing for user interaction or programmatically taking user preferences into account. Finally, dynamic link generation and forwarding schemes may have implications on search results and ranking by existing search engines. Since a wrong choice may be sanctioned by them, implications should be considered and discussed during the evaluation of the visualization.

Implementation

After having covered all entities, tasks, and functions, their composition in an intended solution needs to be realized. This will follow an iterative approach which looks at different stages of cooperation. Starting from two self-contained approaches, in which all functions are joined in a single entity, the development will iterate towards two cooperative approaches, in which the tasks and functions are distributed and shared across multiple entities in an efficient manner.

1. The first self-contained approach will run on a single web server. It needs no interaction with other entities on the internet to function. Its objective is to provide consumers a valuable set of links to resources that are related to their own contents.

2. The second self-contained approach will run on the client side. This approach also needs no interaction with other entities on the internet to function but has a different objective than the first. The generation of valuable links is based on a single user's local knowledge and behavior; it is expected to produce more user-oriented results.

3. In the third approach, the discussed tasks and functions will be provided as web services by intermediaries who operate independently from information providers or consumers. Their services can be utilized by both, web servers and clients. Business models could be driving motivating factors in shaping these services.

4. The last stage of implementation will follow a P2P approach to integrate entities as nodes in a P2P network. The nodes will share data and tasks amongst each other and can be represented by information providers, clients, or intermediaries likewise. The motivation of participants will have to be evaluated. They may be driven by social or egoistic factors, but also by business models.

4.2 Proof of Work for Link Generation

The solution idea, proposed in the previous section, is a high-level concept and needs to be refined and broken down into smaller sub-tasks. The intended distributed link generating and forwarding system may be achieved at a later stage and subsequently brought to the market as a fully developed web service. Breaking the concept down means to focus on certain aspects at the beginning. The first thing to examine will be the process of link generation.

It will be attempted to achieve this by automatically generating links between articles of the freely available Wikipedia article database. Later, it is intended to extend the resulting proof of work by including other web-based content.

Wikipedia serves as an ideal example for the development and application of a proof of work, for various reasons. It provides a well-structured and comparable set of presumedly high-quality text documents. Unified crawling and parsing processes can be applied. Articles already contain links to each other, which are manually maintained by the authors and can be compared with artificially generated links. The same applies to hierarchical links and topic clusters. Most articles also exist in multiple languages, allowing to compare results for equal topics in different languages. In addition, steps and tools for data preprocessing with the aim of text analysis have already been presented in [31] and can be adapted to the purpose of this work.

In summary, Wikipedia provides articles that are heavily interconnected with each other. It is also an open, freely available, and accessible source for anyone. Even though the quality of content and links cannot be assured for all articles, existing guidelines [32] define how links are supposed to be used and are controlled by the community. Therefore, Wikipedia, its articles and internal linking can serve as a reference for automated link generation in this research project. If satisfying results are achieved the proof of work can be extended to larger parts of the web. At that point, other aspects of the proposed research project, such as crawling, ranking, and distributed provisioning will be considered.

5 Conceptual Thoughts on Link Generation

5.1 Link Generation

For the following conceptual thoughts on link generation, a given, comprehensive mapping of keywords and its best-matching information sources' URLs

(term-link-map), as well as a well-developed co-occurrence graph will be presumed.

For a certain reference document, it should be defined how many and which words will be enriched with one or multiple hyperlinks from the term-link-map. Additionally, links may not only be anchored at single words of a text but may also be appended to whole sentences or paragraphs. Approaches for this may range from creating links for every word, name, or substantive that has a matching term on the term-link-map, to a fixed ratio of links per sentence or paragraph.

Overlinking, e. g. setting links for every word obviously seems to be a too excessive approach and most likely results in a useless overload of information. Limiting numbers of links by fixed ratios, on the other hand, will not account for important keywords and relevant sections of a document when a predefined threshold is reached. Therefore, appropriate measures have to be developed in order to find the relevant terms that should be enriched with hyperlinks.

Utilizing text centroids or other text representations could be an approach to find ideal link anchors in a reference document. Since centroid terms may represent whole text bodies, paragraphs or sentences, they can be used to determine a best fitting link from the term-link-map. This can be achieved by measuring distances between words and centroid terms on the co-occurrence graph: choosing the most distant words of a certain text body, paragraph or sentence will allow broader coverage of unrelated, distant topics and creating a deeper interconnection of web content; selecting the closest distances between centroid terms and terms of the term-link-map allows creating links to documents which potentially specialize on a topic which is only marginally discussed in the first document.

While the term-link-map allows creating links to alternative sources for the same topic, the co-occurrence graph allows finding terms that are closely connected with a given centroid term. One additional use case of this could be the generation of link clusters that present links to closely related topics. An example cluster based on a document that revolves around the term ”Car” could look similar to the following outline:

[Car, Driver License, Car Dealer, Opel, Truck,
Formula 1, Tuning, Traffic Regulations]

Another approach is to generate link trees by referring to holonyms (upward-pointing generalizations) and meronyms (downward-pointing specializations) of a document's centroid term. An example tree based on a document that revolves around the term "Car" could look similar to the following outline:

$$\begin{array}{lllll} \text{Vehicle} & \leftarrow & \underline{\text{Car}} & \rightarrow \text{ Engine} & \rightarrow \text{ Piston} \\ & & & \rightarrow \text{ Brakes} & \rightarrow \ \ldots \\ & & & \rightarrow \ \ldots & \end{array}$$

5.2 Further Aspects

The approaches described aim to generate a fixed set of links for a given document. To achieve a more user-oriented service, it should be considered to generate links individually, based on a user's knowledge and preferences. User knowledge can be represented by individually generated local co-occurrence graphs and replace the previously assumed, generic and well-developed graphs. Examining the divergence between the two methods gives additional opportunity to study the effects of individual preferences on the selection and order of content and the appearance of filter bubbles.

Last, but not least, link building should not be seen as a one-off task: new sources emerge, existing sources disappear. The associations with terms may also change over time. This makes it necessary to re-evaluate links and to check their relevance and quality over time. Finally, links may have to be removed or "forgotten".

5.3 Quality of Links

In order to determine the quality of artificially generated links, appropriate measures must be developed. To this end, it should be examined whether existing links in Wikipedia articles can be accepted as a reference, as described in Section 4.2, or dedicated quality measures need to be developed. To achieve this, the Wikipedia link building scheme should be examined more closely and whether the linking follows certain rules or policies. Further literature should also be evaluated with the aim of finding existing research in this area.

6 Summary

This conceptual paper presented the intention of the author to commit research in the area of natural language processing and information retrieval. It has introduced the latest state of the art of several related research topics and illustrated how they could be applied and combined in a new fashion in order to yield the desired goals of strengthening interconnectivity and liberalizing the web. Subsequent research activities in dedicated disciplines will be building upon this, dealing with specific questions, such as the process of link generation.

As discussed in the introduction, the web is constantly changing and a few actors have gained advantages over the vast majority of participants. Eliminating this imbalance is the major driver of the author's motivation and, therefore, primary attention goes to the liberalization of the WWW. This could possibly be achieved through the solution outline and proof of work drafted in chapter four. Its final shape will crystallize in the course of the research project.

Chapter three has shown the difficulties of measuring the change in the nature and source of links over time. Therefore, the analysis of the hypothesis of declining interconnectivity will be deferred, until after more relevant matters have been resolved. In addition, the promising benefits that may potentially be derived from solving the problems presented as the author's motivation would most likely outweigh the rejection of the hypothesis. It is also proposed that the evolution of hyperlinks could be analyzed in a detached empirical study.

The next steps following this introductory paper will be to follow-up the proposals made in chapter four and five. This includes developing a proof of work for link generation and conducting experiments to determine the most efficient link generation scheme. Measures and experiments need to be defined at the beginning. To do this, qualitative references need to be elaborated, e. g. by running empirical surveys or evaluating secondary data. The data preparation and experiments will then be carried out in an experimental environment. Finally, the results are going to be reviewed and conclusions will be drawn.

References

[1] "The birth of the Web", CERN, 2019. [Online]. Available: `https://home.cern/science/computing/birth-web`. [Accessed 11 January 2019].

[2] M. Gray, "Web Growth Summary", 1996. [Online]. Available: http://www.mit.edu/people/mkgray/net/web-growth-summary.html. [Accessed 11 January 2019].

[3] M. Nottingham, "Web Linking", Internet Engineering Task Force (IETF), October 2017. [Online]. Available: https://tools.ietf.org/html/rfc8288. [Accessed 19 February 2019].

[4] "Total number of Websites", Internet Live Stats, January 2019. [Online]. Available: http://www.internetlivestats.com/total-number-of-websites/. [Accessed 11 January 2019].

[5] J. Kleinberg, "Authoritative sources in a hyperlinked environment", *Journal of the ACM. 46*, Nr. 5, pp. 604–632, 1999.

[6] L. Page, "Method for node ranking in a linked database". USA Patent US6285999, 10 January 1997.

[7] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", in: *Seventh International World-Wide Web Conference (WWW 1998)*, Brisbane, Australia, 1998.

[8] J. Leskovec and A. Krevl, "Stanford Large Network Dataset Collection", June 2014. [Online]. Available: http://snap.stanford.edu/data/index.html#web. [Accessed 5 March 2019].

[9] R. Meusel, S. Vigna and O. Lehm, "The Graph Structure in the Web – Analyzed on Different Aggregation Levels", *J. Web Science*, pp. 33–47, 2015.

[10] S. Nagel, "Host- and Domain-Level Web Graphs Nov/Dec/Jan 2018 – 2019", Common Crawl, 20 February 2019. [Online]. Available: http://commoncrawl.org/2019/02/host-and-domain-level-web-graphs-nov-dec-2018-jan-2019/. [Accessed 5 March 2019].

[11] K. Dooley, Designing Large-scale LANs - Page 31, O'Reilly, 2002.

[12] Amazon, "Summary of the Amazon S3 Service Disruption in the Northern Virginia (US-EAST-1) Region", Amazon, February 2017. [Online]. Available: https://aws.amazon.com/message/41926/. [Accessed 19 February 2019].

[13] E. Pariser, The Filter Bubble: What the Internet Is Hiding from You., New York: Penguin Press, 2011.

[14] BBC, "Google turns off China censorship warning", 7 January 2013. [Online]. Available: https://www.bbc.com/news/technology-20932072. [Accessed 5 March 2019].

[15] European Commission, "Antitrust: Commission fines Google 4.34 billion for illegal practices regarding Android mobile devices to strengthen dominance of Google's search engine", 18 July 2018. [Online]. Available: http://europa.eu/rapid/press-release_IP-18-4581_en.htm. [Accessed 23 February 2019].

[16] StatCounter, "Worldwide desktop market share of leading search engines from January 2010 to October 2018", May 2019. [Online]. Available: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/. [Accessed 20 February 2019].

[17] M. Kubek and H. Unger, "The WebEngine – A Fully Integrated, Decentralised Web Search Engine", November 2018.

[18] B. J. Jansen, D. L. Booth and A. Spink, "Determining the user intent of web search engine queries", *WWW '07 Proceedings of the 16th international conference on World Wide Web*, pp. 1149–1150, 08-12 May 2007.

[19] M. Kubek and H. Unger, "Centroid Terms as Text Representatives", in: *Proceedings of the 2016 ACM Symposium on Document Engineering*, pp. 99–102, September 2016.

[20] M. Kubek and H. Unger, "On Evolving Text Centroids", *Recent Advances in Information and Communication Technology 2018*, pp. 75–82, 27 June 2018.

[21] K. C. G. C. J. D. Tomas Mikolov, "Efficient Estimation of Word Representations in Vector Space", *CoRR*, no. abs/1301.3781, 16 January 2013.

[22] T. M. Quoc V. Le, "Distributed Representations of Sentences and Documents", *CoRR*, 16 May 2014.

[23] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system", 31 October 2008.

[24] S. Sodsee, *Placing Files on the Nodes of Peer-to-Peer Systems*, Hagen, 2011.

[25] I. Stoica, R. Morris, D. Liben-Nowell, D. R. Karger, M. F. Kaashoek, F. Dabek and H. Balakrishnan, "Chord: a scalable peer-to-peer lookup protocol for internet applications", *IEEE/ACM Transactions on Networking (TON)*, pp. 17–32, February 2013.

[26] G. Sakaryan, *A Content-Oriented Approach to Topology Evolution and Search in Peer-to-Peer Systems*, Rostock, 2004.

[27] M. Kubek, *Dezentrale, kontextbasierte Steuerung der Suche im Internet*, Hagen, 2012.

[28] eMarketer, "Number of social media users worldwide from 2010 to 2021 (in billions)," July 2017. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. [Accessed 03 July 2019].

[29] Statista, "Social network penetration worldwide from 2014 to 2021", July 2017. [Online]. Available: https://www.statista.com/statistics/260811/social-network-penetration-worldwide/. [Accessed 03 July 2019].

[30] "Common Crawl," The Common Crawl Foundation, May 2019. [Online]. Available: http://commoncrawl.org/the-data/get-started/. [Accessed 25 June 2019].

[31] O. Davydova, "Text Preprocessing in Python: Steps, Tools, and Examples", Data Monsters, 15 October 2018. [Online]. Available: https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908. [Accessed 01 July 2019].

[32] I. Wikimedia Foundation, "Wikipedia:Manual of Style/Linking", 20 June 2019. [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking. [Accessed 25 June 2019].

Balancing Data Access Frequencies in P2P Systems with Error-prone Communication

Dimitri Samorukov

Chair of Communication Networks, FernUniversität in Hagen, Germany

Abstract: Distributed applications are run decentrally on peers who volunteer their resources to the application. Various distributed applications require a data set for their function. This data set can be very large and must therefore be distributed among several peers. Since the peers can change their available resources at any time or even depart, parts or all of the data set must remain mobile. Here a solution is presented which allows to move the data sets between the peers. The unreliability of the communication connections and the requirement for parallel, non-blocking migration are taken into account.

1 Motivation

Various currently existing applications are distributed and require an existing peer community for their function. Worth mentioning are distributed web searches, social networks, linked data initiatives and digital currencies.

The distributed web searches [4, 9, 12, 13] provide an alternative to the commercial and centralized approaches of web search applications. Social networks in their distributed form [5] allow the user full control over his data. The linked data initiatives [10] based on the Symantic Web approach form a globally distributed data set.

Digital currencies have long relied on distributed, replicated databases. Each participating peer has a full copy of the entire data set [17]. However, this procedure has proven to be very disadvantageous, as only a few peers are willing and able to hold the data set, which is several hundred gigabytes in size. Few peers have got a big influence and this increased the risk of malicious attacks. Therefore, approaches were developed [18] to distribute the one large data set replicated over several peers in smaller, interconnected units across the peer community. Thus, these applications have one thing in common: they need a data set that is distributed decentrally among the peers of the community.

The distributed data set (also administrative structure) poses special challenges for the application, in detail it is:

- higher susceptibility to data loss due to random peer churn
- increasing response times of the application compared to a centralized solution
- responsibility for resources, the application is now faced with the task of collecting sufficient resources from the community itself, instead of assuming their availability as in the centralized approach.

The decentralized applications have better privacy features, the solution of upper problems is crucial for the better competitiveness of the decentralized approaches against centralized versions.

In [19] it was shown that a suitable placement of the management structure on the peers can significantly accelerate the execution of operations on the structure. However, this requires a reliable and non-destructive movement of the administrative structure between the peers.

This paper provides the answer to the question: how to move the administrative structure between peers a) without destroying it b) without peer overload in case of unreliable communication between peers.

2 System Model

The system model consists of a graph of peers $G_p = (K_p, C_p)$ $C_p = K_p \times K_p$ which build a full connected graph with undirected connections. A peer $k_p \in K_p$ is defined by provided and currently free resources $k_p = (R_p^b, R_p)$, where R_p^b are provided resources and R_p currently free resources. Decision about provided resources is felt by *operator* of the peer. We assume that there is no spontaneous churn among the peers. Prior leaving the network, the *operator* reduces the provided resources R_p^b down to 0. Communication over connections c_p is unsecured, so any packets may be lost with probability of p^{lost}, it is a property of each connection c_p. This is independent of the length of the transmitted packet and the transmission direction. c_p has a latency of null and is able to transmit N_p packets per time step Δt.

Real-World applications setup an additional graph, so-called overlay graph, which is used for routing purposes. This is defined here as $G_p^o = (K_p, C_p^o)$ with

$C_p^o \subseteq K_p \times K_p$,$G_p^o$ is undirected. The neighbourhood $K_{nb} \in K_p$ of peer k_p is defined over the connections C_p^o.

The graph of administrative structure $G_v = (K_v, C_v)$ where $C_v \subseteq K_v \times K_v$ has an arbitrary topology. Connections $c_v \in C_v$ are undirected. k_v is defined by required resources $k_v = (R_v)$, with the same unit as k_p parameters R_p and R_p^b.

The embedding of graph G_v into G_p is defined by a mapping $\Phi : K_v \to K_p$. It is implicitly assumed that if $(k_v^1, k_v^2) \in C_v$, then $\Phi(k_v^2)$ is 'known' in k_v^1. Each node $k_v \in K_v$ and $k_p \in K_p$ holds a list of all it's neighbours. Knowing $\Phi(k_v^2)$ means that k_v^1 also contains a list ($mapping - list$) of all $\phi(k_v^2)$ with $(k_v^1, k_v^1) \in C_v$. Local nodes of a peer k_p are denoted as a set $K_v^{mg} = \{k_v \in K_v | \forall k_v : \Phi(k_v) = k_p\}$. Φ is only valid if node k_v is always assigned to exactly one peer, more formal: $\forall k_v \in C_v : (\exists! k_p \in K_p : \Phi(k_v) = k_p)$.

We define a peer to be overloaded if $R_p < 0$, with $R_p = R_p^b - \sum_{K_v^{mg}}^{k_v} R_v(k_v)$. A peer gets overloaded either by reduced resources by *operator* or additionally received nodes k_v. If overloaded it randomly selects n local hold nodes $K_l \subseteq K_v^{mg}$ so that $g = \sum_{K_l}^{k_v} R_v, R_p + g \geq 0$ and then migrates them to randomly selected peer $k_p \in K_p$.

Migration ($migrate(k_v, k_p)$) changes the embedding of G_v into G_p. It assigns the node k_v to new node k_p. This changes the current mapping from Φ to new mapping Φ'. This leads to updating of $mapping - list$ of all neighbours of k_v to new peer k_p. Due to insecure communication the migration can fail.

The failure of ($migrate(k_v, k_p)$) is detected when the time T_{tr} is reached. After this, the operation is repeated with another, randomly selected peer of the neighbourhood K_{nb}.

The transmission time over c_p is modelled by a FIFO memory of infinite size on the local peer and a sender. The shipper takes Δt N_p packets from the buffer and sends them to the neighbouring peer. A newly arriving packet has to wait $t_w = BufferLength\ / N_p + 1$ for its arrival at the target peer, with $BufferLength$ as number of entries currently contained in the buffer. If now $T_{tr} < t_w^1 + t_w^2$, then the operation will fail, with t_w^1, t_w^2 as delay times for each direction.

3 Problem

The *operator* expects his peer to be in an overloaded state as rarely as possible. For this to happen, an overload must be resolved as quickly as possible. On

the other hand, the *user* expects the administrative structure to be complete and accessible at all times.

A naive implementation of the operation $migrate(k_v, k_p)$ performed by a peer would transfer a local node k_v to the new peer in a serialized way. This means that after the transfer from the source peer k_p, k_v notifies its neighbours to adjust their $mapping - list$. However, the entries in k_v's $mapping - list$ must remain current after arrival on a new peer. This implies that neighbours cannot migrate until the moved node has arrived and the mapping-lists of the neighbours have been adjusted. However, communication over the c_p connections is insecure due to $p^{lost} > 0$, new packet transfers are a common means. But, they can lead to unwanted k_v copies on several peers k_p. The insecure communication together with the compulsion of serializing the transmission requires a deeper exploration.

The addressed problem here is a migration operation which fulfils following requirements:

a) leaves the graph G_v intact (correctness)

b) creates another valid mapping Φ (duplicate avoidance)

c) allows a parallel migration of all nodes $k_v \in K_v$ without locking procedures (locking freedom)

The main conditions 'correctness' and 'duplicate avoidance' ensure that no nodes or connections are lost or duplicated through migration to a peer. In particular, 'duplicate avoidance' reduces the burden on the peers involved. Without ongoing migrations, the duplicates should be cleaned up in a finite time. The constraint 'locking freedom' ensures the satisfaction of the *operator* needs on fast resources usage adaptation. Locking should be completely avoided during the migration of nodes.

4 Proposed Solution

The solution bases on *forwarding pointer* approach [3]. The basic concept is that each node of the graph leaves a path in the network that allows the migrated node to be found on other peers. However, due to unreliable communication, additional actions must be developed to detect and remove duplicated nodes.

4.1 Path of a Migrated Node k_v

For this purpose, each peer is extended by a buffer called *forwarding pointer cache* C_{fpc}. Here each migrated node k_v is entered after successful migration. Thus *forwarding pointer cache* is a set of tuples $C_{fpc} = \{(k_v, c_p^o), \ldots\}$ with a unique entry for a k_v. The size of *forwarding pointer cache* C_{fpc} has an upper limit S_{fpc}. If the number of entries exceeds the S_{fpc}, so a cache clean-up strategy CT selects an entry and removes it from C_{fpc}. Examples for CT are FIFO (first-in-first-out), LRU (least recently used), LRD (Least Reference Density, [8]).

To reach a neighbour k_v-node holds it's k_p in the local $mapping - list$. After migration of the neighbour, node k_v needs to follow the *path* left by the migrated neighbour over the peers k_p and their *forwarding pointer caches* C_{fpc}. This path is defined by a ordered set $path \subseteq C_p^o$. Thus the $path - length$ defines the number of jumps between peers to reach the neighbour node of k_v. If the *path* is empty, then the neighbour was not migrated, and the entry from $mapping - list$ is current. The *path* is called broken if the neighbour k_v not contained on the last k_p of the *path*. To each $c_v \in C_v$ are two *path* – properties assigned, one for each direction.

4.2 Authentication and Garbage Collection

Due to unreliable communication over connections c_v with p^{lost} lost-packet rate, possible resending of nodes k_v to other k_p can create several copies of migrated node k_v. Therefore an authentication state $AuthState$ is assigned to a node k_v with values 'authenticated' and 'not authenticated'. Immediately after arriving at the target peer, the status $AuthState$ is 'not authenticated'. The authentication process is executed by the node k_v as long as $AuthState =$ 'not authenticated'. It is repeated with period of T_{Auth} as long as $AuthState =$ 'not authenticated'. Authentication process consists of following steps:

a) ask each neighbour to follow the *path* and check if the *path* leads to k_p of asking k_v

b) when each of the neighbours of k_v have reached k_v over their path, k_v sets it's $AuthState =$ 'authenticated'

c) during authentication process, if migration was successful, all neighbours update their $mapping - list$ to set $path - length = 0$.

To increase free resources of each k_p unauthenticated nodes k_p should be detected and removed. Deleting these unauthenticated nodes is a stochastic, time

based process called garbage collection. It bases on increasing probability for deletion of a node with $AuthState$ ='not authenticated' with time t. Directly after the arrival on the peer k_p at the time t_A, the nodes k_v for the period $T_{VirtAuth}$ are regarded as virtually authenticated. During this time they can authenticate themselves undisturbed. After $T_{VirtAuth}$ period at time $t_d^1 = t_A + T_{VirtAuth}$ the probability p_D of deletion increases until it gets 1 at time $t_d^2 = t_d^1 + N_f \cdot T_{VirtAuth}$, with $N_f \geq 1$ as an adjustment factor. Thus we can specify the probability p_D as function of time t for node k_v, by:

$$p_D(t) = \begin{cases} \frac{(t - t_d^1)}{t_d^2 - t_d^1} & t \geq t_d^1 \wedge t \leq t_d^2 \wedge AuthState(k_v) = \text{not authenticated} \\ 1 & t \geq t_d^2 \wedge AuthState(k_v) = \text{not authenticated} \\ 0 & \text{else} \end{cases} \quad (1)$$

Time period $T_{VirtAuth}$ should be longer that the authentication process period T_{Auth}. Thus we define it by $T_{VirtAuth} = N_b \cdot T_{Auth}$ with $N_b \geq 1$. T_{Auth}, N_b and N_f are peer properties.

4.3 Communication Protocol Details

Based on previous information, protocol details are given here. We distinguish between services and protocols. Services are provided by the peer and node. A protocol is used for communication between two equivalent participants (peers or nodes) using the services. Peer k_p provides the following services: path-resolving communication, node migration. Nodes k_v provide the authentication service.

Path-resolving communication service is used to deliver *payload* to node k_v, following the *path* between peers. The following protocol messages are defined for this purpose: $SendMessage(payload, k_v)$, $ResponseMigratedTo(k_p)$, *ResponseOK()*. The service is called with target node k_v , last known peer k_p from $mapping - list$ and corresponding *payload*. Then the message $SendMessage(payload, k_v)$ is sent to the peer k_p. If the node k_v was contained on the peer, the transmission is confirmed with *ResponseOK()*. If the node was not contained and there was no corresponding entry in C_{fpc} on targeted k_p, no response will be sent. If, however, an entry for k_v with a new peer is found, the response is $ResponseMigratedTo(k_p)$ with the corresponding peer as next node in *path*. The calling peer now sends the $SendMessage(payload, k_v)$ to the new target peer. Each reception of $ResponseMigratedTo(k_p)$ updates the $mapping - list$ of k_v, which shortens the path by one entry.

The node migration service consists of the following protocol messages: *SendMigratingNode*(k_v), *MigratedNodeAccepted*(). In case of peer overload a local node k_v is selected and sent using *SendMigratingNode*(k_v) to one of the neighboring peers. This must confirm the reception with *MigratedNodeAccepted*(). The reception of *MigratedNodeAccepted*() leads to new entry within C_{fpc} of the sender peer. If no confirmation *MigratedNodeAccepted*() was received within time T_{tr}, the procedure is repeated with another, randomly selected neighbour peer.

The authentication service of the node k_v consists of the following protocol messages: *AuthenticateMe*(), *Authenticated*(). Once started, k_v periodically with T_{Auth} sends the message *AuthenticateMe*() to each of its neighbours. It uses the peer's path-resolving communication service. The recipient of the message confirms receipt with an *Authenticated*() message. The service is completed when message *Authenticated*() is received from each neighbour k_v.

5 Evaluations

We look at how far the demands for correctness duplicate avoidance and locking-freedom from the chapter 3 were fulfilled.

5.1 Theoretical Reflection

It's obvious that the locking-freedom requirement has been met. By leaving the *path* when migrating, the migrating node does not have to schedule any wait times for updating the $mapping - list$ of neighbouring nodes either before or after the migrating.

The requirement for *correctness* implies that neither nodes nor connections are lost. The nodes can be deleted by garbage collection procedure. The c_v connections are lost by the *path* break.

The cause of *path* break is the overflow of the *forwarding pointer cache* C_{fpc} due to its limited size S_{fpc}. The clean-up-strategy CT has a significant influence on this. The entries in C_{fpc} should only be deleted if there are no more queries for them. Immediately after the migration of a node k_v, all neighbours of k_v will follow the *path* and determine the new peer. Only when this first wave has decreased, the entry in C_{fpc} can be deleted. The duration of this first wave depends on the number of connections of the migrated node. Therefore, the entry in C_{fpc} should exist as long as it is queried. Therefore, the clean-up strategies LRU (least recently used) and LRD (least reference density) seem to be the most suitable.

The reason for deleting nodes is that authentication did not take place and the garbage collection process removed it. The reason for this is a) the node k_v is a duplicate b) the node k_v is not a duplicate, but authentication is not yet complete. The reason for the latter is the long duration t_a of authentication ($t_a > T_{VirtAuth}$). The duration t_a is proportional to the own $path - length$, the $path - length$ to the neighbours and the packet loss rate p_{lost}. However, the $path - length$ only increases when migrations of the own and neighbour nodes k_v take place. Thus, in a network with many migrations, more node losses occur, especially if the ratio $T_{VirtAuth}/T_{Auth} \approx 1$.

We now consider the demand for duplicate avoidance. In an unsafe communication duplicates of transmitted nodes k_v can occur. A migration consists of several migration attempts. A successful migration attempt is sending the message *SendMigratingNode*(k_v) and receiving the corresponding response *MigratedNodeAccepted*(). Every message can be lost with probability p^{lost}. The probability for a successful attempt is $p_{succ} = (1 - p^{lost})^2$. For the creation of a duplicates only the answer *MigratedNodeAccepted*() has to be lost, the probability is $p_{doubler} = (1 - p^{lost}) \cdot p^{lost}$. Now we ask for the number n_{tries} of necessary migration attempts for successful node migration with the limit probability p_{lim}. $p_{lim} = 1 - (1 - p_{succ})^{n_{tries}}$ applies here. After conversion $n_{tries} = log(1 - p_{lim})/log(1 - p_{succ})$. This gives us the expected number of duplicates generated per migration: $N_{doubler} = n_{tries} \cdot p_{doubler}$. If the original peer has more than one neighbour, the $N_{doubler}$ will be evenly distributed among the neighbours.

However, this modelling does not take into account the failed migrations due to traffic volume and thus the retries after T_{tr} and $path$ breaks. Therefore, p_{succ} will be lower in the real world, the more packets are transferred over c_p connections.

The table 1 shows the above values for different packet failure probabilities p_{lost} and limit probability $p_{lim} = 0.999$. Thus, 684 duplicates are generated per migration at an average packet loss rate of 99% and the migration is completed after 69074 attempts. The result of this is that even at a high loss rate of 99%, the migrations do not last indefinitely and the duplicate node generation is only nominal at high loss rates. And here these nodes can be recognized and deleted after $N_f \cdot T_{VirtAuth}$ time with 100% probability. See equation (1). After starting migrations until the time $T_{VirtAuth}$ the number of duplicates increases. However, as of $T_{VirtAuth}$ they will be deleted with the garbage collection process and their number will increase less. Thus the number of duplicates in the system

depends on the rate p_{lost}, time $T_{VirtAuth}$ and the number of migrations taking place. If there are no migrations, the number of duplicates will run against 0.

Table 1: System behavior when duplicating nodes

p^{lost}	n_{tries}	$p_{doubler}$	$N_{doubler}$
0.1	5	0.09	1
0.2	7	0.16	2
0.3	10	0.21	3
0.4	16	0.24	4
0.5	24	0.25	6
0.6	40	0.24	10
0.7	74	0.21	16
0.8	170	0.16	27
0.9	688	0.09	62
0.99	69074	0.0099	684

5.2 Simulative Results

The upper predictions are verified in a simulation. For this a step response of the system to a reduction of the provided resources R_p^b is simulated. The graph G_p is completely connected and consists of 100 peers with $R_p^b = \infty$. The parameter N_p of the connections c_p is $\forall c_p \; N_p = 10$. p^{lost} has also the same value for all c_p. Also values of the parameters T_{Auth}, N_b and N_f are valid for all peers k_p. The graph G_v is fully connected and consists of three nodes k_v. The required resources for the nodes k_v are set to $R_v = 1$. At the beginning of the simulation, the graph G_v is assigned to exactly one *source peer* k_p and the resources R_p^b of the peer are reduced. k_p now tries to move local nodes k_p to the neighborhood K_{nb} with the function $migrate(k_v, k_p)$. In this case $G_p^o = G_p$ applies. For each time step Δt, each peer k_v is assigned computing time once, so that it can process its incoming messages and generate outgoing messages. The step response is considered completed when a) there are no pending messages and b) all nodes k_v are authenticated. The step response is considered *successful* if the original graph G_v was not changed after completion.

The migration of the overloaded node is repeated by T_{tr} Δt steps, which is set here to $T_{tr} = 3$. The garbage collection has following settings. With $T_{Auth} = 15$ authentication is repeated each 15 Δt steps. $N_b = 3$ leads to $T_{VirtAuth} = 45$,

thus after 45 Δt steps an unauthenticated k_v may be deleted. With $N_f = 1000$ unauthenticated nodes will be deleted after 45000 Δt steps. The size S_{fpc} of the *forwarding pointer cache* is set to $S_{fpc} = 3$. Therefore, no path break is expected.

The table 2 shows the sequence of a step response in the simulation in tabular form for $p^{lost} = 0$ for all c_v and $R_p^b = 2$ for *source peer*. This means that exactly one node is migrated. After 6 Δt steps, the step response is completed. Here $\mathbf{N_{K_v}}$ stands for the number of nodes k_v on all peers k_p, $\mathbf{N_{K_v}^{NotAuth}}$ for the number of non-authenticated nodes, $\mathbf{N_{K_v}^{Mgr}}$ for the number of nodes that are currently being migrated, $\mathbf{N^{Msg}}$ for the number of messages that are being transmitted but not yet processed.

Table 2: Steps of a single simulation step

$\mathbf{\Delta t}$	0	1	2	3	4	5
$\mathbf{N_{K_v}}$	3	4	3	3	3	3
$\mathbf{N_{K_v}^{NotAuth}}$	0	1	1	1	1	0
$\mathbf{N_{K_v}^{Mgr}}$	0	1	0	0	0	0
$\mathbf{N^{Msg}}$	0	1	2	2	2	0

We repeat the evaluation of the step response 1000 times and trace the maximum $\mathbf{N_{K_v}}$ that occurred, the average number of migration attempts $\mathbf{n_{tries}}$ and the average migration duration in steps $\mathbf{n \cdot \Delta t}$ and the number of deleted duplicate nodes $\mathbf{N_{doubler}}$. Only *successful* step responses are considered. R_{succ} is the success rate of the step responses. These attempts are performed for $R_p^b = 2$, $R_p^b = 1$ and $R_p^b = 0$ of *source peer* k_p and different p_{lost} rates. The results are shown in the tables 3, 4 and 5.

If we compare these results with the expectations from table 1, we notice a discrepancy at $\mathbf{n_{tries}}$ and $\mathbf{N_{doubler}}$. However, this can be explained by the mentioned inaccuracy of the model at retransmissions (time out T_{tr}) caused by N_p limitations of connections c_p.

6 Related Work

The current work has strong relation to the area of mobile agents (MA). Nodes k_v can be regarded as a light weight mobile agent, with a few, unchanging com-

Table 3: Success rate of migrations with one migrating node k_v, $R_p^b = 2$

$\mathbf{p^{lost}}$	$\mathbf{N_{K_v}}$	$\mathbf{N_{doubler}}$	$\mathbf{n_{tries}}$	$\mathbf{R_{succ}}$	$\mathbf{n \cdot \Delta t}$
0	4	0	0	1	5
0.1	7	3	4	1	4969
0.2	8	4	6	1	8995
0.3	9	5	7	1	13975
0.4	11	7	13	1	17123
0.5	12	8	25	1	22305
0.6	14	10	36	1	27927
0.7	22	18	69	1	32657
0.8	33	29	134	0.992	39646
0.9	51	47	612	0.196	44017

Table 4: Success rate of migrations with two migrating nodes k_v, $R_p^b = 1$

$\mathbf{p^{lost}}$	$\mathbf{N_{K_v}}$	$\mathbf{N_{doubler}}$	$\mathbf{n_{tries}}$	$\mathbf{R_{succ}}$	$\mathbf{n \cdot \Delta t}$
0	5	0	0	1	5
0.1	7	3	4	0.998	7646
0.2	8	4	7	0.993	16764
0.3	10	6	12	0.994	22025
0.4	12	8	16	0.996	28330
0.5	15	11	26	0.993	34788
0.6	22	18	42	0.991	37320
0.7	29	25	87	0.989	41171
0.8	35	31	189	0.965	43607
0.9	72	68	399	0.048	44393

Table 5: Success rate of migrations with three migrating nodes k_v, $R_p^b = 0$

$\mathbf{p^{lost}}$	$\mathbf{N_{K_v}}$	$\mathbf{N_{doubler}}$	$\mathbf{n_{tries}}$	$\mathbf{R_{succ}}$	$\mathbf{n \cdot \Delta t}$
0	6	0	1	1	4
0.1	9	4	4	0.983	12111
0.2	10	5	7	0.984	20895
0.3	11	7	11	0.965	30451
0.4	16	12	16	0.973	34701
0.5	20	16	36	0.965	39064
0.6	28	24	44	0.968	42034
0.7	34	30	87	0.972	44085
0.8	46	42	194	0.938	44800
0.9	51	47	321	0.012	41893

munication partners. The mobile agent research provides several solutions for locating a specific MA and message delivery between MAs.

Choi [3] suggests following classification for locating of MAs: *location server* [24], *forwarding pointer* [16], *brute force* [20]. Wang [22] has chosen to distinguish the MA location by hierarchy-based [22], region-based [21] and home-based methods [22].

The kind of message delivery is another important aspect of each MA system, Choi [3] and Deugo [7] distinguish between *Direct*, *Group*, *Blackboard*, *Mailbox* and *Forwarding* approaches. The *Direct* method allows MA's direct communication. In the *Group* method the message is sent to a set of mobile agents simultaneously, as e.g. a multi- or broadcast [20]. In *Blackboard* the message is placed on a dedicated host, MA can read the message on visiting this host. In the *Mailbox* approach, messages are sent over a Mailbox to a mobile agent. The *Mailbox* can be placed on any peer within the network, it can also be moved across network [2, 14]. It is in responsibility of assigned MA to fetch messages from it's Mailbox. In *Forwarding* method messages are sent to a dedicated MA by following it's migration path. A good overview for this problem is given in [23].

As defined by [3] in *location server* approach there is a central instance, that manages locations of all MA's in the system. In *forwarding pointer* approach each MA leaves on migration his new address, the *broad cast* method determines the location of a specific MA by a broad cast message, sent to all network hosts.

In *Hierarchy-based* architecture defined by Wang [22] exists a central instance, that manages the positions of all MA's within the whole MA system. The managing instance is either a host or a MA. Examples for such systems are [6] and [11]. In [11] the proposed IAgent is either a stationary or mobile agent. A *region based* approach divides the global network into regions, each region contains a dedicated location server. The migration of MA's is either inter-region or extra-region. One example for such system is [21]. *Home-based* location management is another suggested approaches group. To this group belong systems where a dedicated peer manages the position of MA's created by this peer. *Home-based* methods can be divided in *home-proxy* (HP) scheme and *forwarding-proxy* (FP) scheme.

In [16] Moreau developed a fault tolerant, MA location method based on redundant *forwarding pointer*. Each MA remembers N nodes he has visited in the past. After migration this N sites are notified about the new location of the MA. Each node remembers N last positions of the MA and according time stamp. Thus

this solution is very tolerant for loss of nodes. The current position of the MA is stored redundantly on N nodes. As long as one of this nodes exists, the MA can be located.

Ahn [1] extends the solution from Moreau [15] to be more efficient, in terms of communication speed. In [15] a *forwarding pointer* MA location system was developed. Here MA notifies it's previous nodes and a distributed MA location directory, about the new position (see also [16]). Ahn in [1] extends this system, thus the distributed MA location directory is notified only after k migrations. As the location server one of the visited nodes by MA is selected. The newly selected location server stays for k migrations MA's location server. After k migrations a new location server is selected. MA notifies the current location server about each migration. This solution reduces the length of the *forwarding pointer* path.

7 Conclusion and Outlook

The naive solution can avoid the creation of duplicates, but at the expense of serializing the migrations. This problem is solved here. As a side effect, however, duplicates of nodes can arise which have to be detected and removed with additional effort.

Due to the limited size S_{fpc} of the *forwarding pointer cache* C_{fpc} , the path can be broken. Thus the probability of a break increases with the $path - length$. We solve this problem by reducing the path length on authentication.

The resulting duplicates of the nodes are reliably detected and removed. The experimental evaluation showed that on average in more than 90% of all migrations the process was successful within the given N_f limits. Thus the R_p^b resources of the peers are preserved and if no further migrations take place the number of duplicates is reduced automatically.

With this approach a reliable migration of the administrative structure over the peers of the community is possible, while the peers are still able to perform the migration operations independently of each other.

In further steps it is necessary to examine the influence of the length S_{fpc} of the *forwarding pointer cache* and the number of migrations in the network on the stability of $path$. In the experiments it was assumed that $S_{fpc} = \infty$, so the $path$ was never broken.

References

[1] J. Ahn. Decentralized inter-agent message forwarding protocols for mobile agent systems. In *Computational Science and Its Applications–ICCSA 2004*, pages 376–385. Springer, 2004.

[2] J. Cao, L. Zhang, J. Yang, and S. K. Das. A reliable mobile agent communication protocol. In *Distributed Computing Systems, 2004. Proceedings. 24th International Conference on*, pages 468–475. IEEE, 2004.

[3] S. Choi, M. Baik, H. Kim, E. Byun, and H. Choo. A reliable communication protocol for multiregion mobile agent environments. *Parallel and Distributed Systems, IEEE Transactions on*, 21(1):72–85, 2010.

[4] M. Christen. `http://yacy.de/`, 2015.

[5] F. de Asís López-Fuentes and S. Balleza-Gallegos. Comparison of p2p social network architectures. *Research in Computing Science*, 78:43–54, 2014.

[6] J. Desbiens, M. Lavoie, and F. Renaud. Communication and tracking infrastructure of a mobile agent system. In *System Sciences, 1998., Proceedings of the Thirty-First Hawaii International Conference on*, volume 7, pages 54–63. IEEE, 1998.

[7] D. Deugo. Mobile agent messaging models. In *Autonomous Decentralized Systems, 2001. Proceedings. 5th International Symposium on*, pages 278–286. IEEE, 2001.

[8] W. Effelsberg and T. Haerder. Principles of database buffer management. *ACM Transactions on Database Systems (TODS)*, 9(4):560–595, 1984.

[9] FAROO(ltd). `www.faroo.com`. URL, 2015.

[10] O. Görlitz. Distributed query processing for federated rdf data management. 2015.

[11] G. Kastidou, E. Pitoura, and G. Samaras. A scalable hash-based mobile agent location mechanism. In *Distributed Computing Systems, 2003. Proceedings. 23rd International Conference on*, page 472. IEEE, 2003.

[12] M. Kubek and H. Unger. A concept supporting a resilient, fault tolerant and decentralised search. *Autonomous Systems 2017*, 10:20–31, 2017.

[13] M. M. Kubek and H. Unger. Towards a librarian of the web. In *Proceedings of the 2nd International Conference on Communication and Information Processing*, pages 70–78. ACM, 2016.

[14] Z. Lu, D. Liu, G. Zeng, and G. Yang. The communication model of migrating workflow system. In *Agent Computing and Multi-Agent Systems*, pages 496–501. Springer, 2006.

[15] L. Moreau. Distributed directory service and message routing for mobile agents. *Science of Computer Programming*, 39(2):249–272, 2001.

[16] L. Moreau. A fault-tolerant directory service for mobile agents based on forwarding pointers. In *Proceedings of the 2002 ACM symposium on Applied computing*, pages 93–100. ACM, 2002.
[17] S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008.
[18] S. Popov. The tangle. *cit. on*, page 131, 2016.
[19] D. Samorukov. Operation on distributed structures: An optimization approach. In *4th Workshop Complex Structures*, number 4. Forschungsinstitut für Telekommunikation und Kooperation, 2019.
[20] O. Shehory. A scalable agent location mechanism. In *Intelligent Agents VI. Agent Theories, Architectures, and Languages*, pages 162–172. Springer, 1999.
[21] A. D. Stefano and C. Santoro. Locating mobile agents in a wide distributed environment. *Parallel and Distributed Systems, IEEE Transactions on*, 13(8):844–864, 2002.
[22] T. I. Wang and T. Y. Yeh. Development of low cost message delivery path for mobile agent communication. *Journal of Convergence Information Technology*, 5(2):164–173, 2010.
[23] P. T. Wojciechowski. Algorithms for location-independent communication between mobile agents. Technical report, 2001.
[24] M. Xie. *Locating mobile agents through distributed mechanisms*. PhD thesis, University of Ottawa (Canada), 2008.

Thai Word Segmentation by Automatic Ranking Trie with Misspelling Correction

Chalermpol Tapsai[1], Phayung Meesad[1] and Choochart Haruechaiyasak[2]

[1]King Mongkut's University of Technology North Bangkok, Thailand
[2]National Electronics and Computer Technology Center (NECTEC), Thailand

Abstract: Word segmentation is a very important process in Natural Languages Processing (NLP) especially for non-segmentation languages such as Thai, Chinese, and Japanese, which make the word segmentation process more complicated and easily cause errors. Thai word segmentation programs have been developed and improved by many techniques but still have 3 main disadvantages, i.e., dictionary size is quite big, too many excessive parsing tasks on unused words, and no misspelling word handling. In this research, we propose two new techniques, including Ranking Trie and Completed Soundex, that help exclude the unused words from the dictionary, reduce parsing tasks, and correct the misspelling words. The data used in this research are 6,579 text files collected from various sources of data in both online and offline covered 10 content's types, including economy, society, politics, health, education, agriculture, entertainment, sports, technology and IT, and others. The data was divided into 2 sets: 1) Learning set which are 5,879 text files randomly selected and used for word segmentation model developing, 2) Test set, the rest 700 text files used for testing and performance evaluation. The results showed that the performances on word segmentation and misspelling correction are very good with all values of accuracy, precision, recall, and f-measure greater than 0.9.

1 Introduction

For many years, studies on NLP have been conducted in order to help the non-technician users to command and utilize computer by their own languages. To make computers understand a human language, the first step in NLP is Lexical analysis which will split sentences into words and defined types. Thai language

is a non-segmentation natural language which all words are continuously written in sentences without any delimiters. This written style make word segmentation process more complicated and easily cause errors. According to many studies, Thai word segmentation programs were developed and improved by many techniques until now, bring about the most widely used program names âœLextoâ. By using a Trie structure and longest match technique, Lexto works well but still have 3 main disadvantages, including dictionary size is quite big, too many excessive parsing tasks on unused words, and no misspelling word handling. For this reason, the researchers are interested in developing an algorithm of Thai word segmentation that can be used to solve these disadvantages. The rest contents of this article are divided into 6 parts: Related works, Research concept, Model development, Experiment and performance evaluation, Experimental Results, and Conclusion.

2 Related Work

2.1 Thai Word Segmentation

Many studies focus on Thai word segmentation have been researched and developed for more than 30 years. These studies were conducted by various techniques which can be divided into 4 types, including rule-based word segmentation (RB-WS), dictionary-based word segmentation (DB-WS), and machine learning-based word segmentation (MLB-WS). RB-WS is the first phase of Thai word segmentation algorithm that was used to analyze the experimental texts to define the boundary of syllables which is an elementary component of a word [1, 2]. By using linguistic spelling principles which related to the characteristics of each type of Thai alphabets, the specific rules were set up to identify the front boundary and the rear boundary of each syllable. Although this method can identify the syllable's boundaries with very high accuracy, most of Thai words consist of more than one syllable. Therefore, it doesn't work well for word segmentation. DB-WS was firstly used for Thai word segmentation by Pooworawan in 1986 [3]. This algorithm was designed for determining the boundaries of syllables by parsing the input text with a dictionary of Thai syllables with the longest match strategy. Unfortunately, although this method has a high degree of correctness in syllable scoping, it doesn't design for word segmentation. This problem is later improved by Raruenrom [4] who changed the content of the dictionary from syllables to words and reduce the time for parsing by using the Trie structure [5] to organize words in the dictionary. At the later time, DB-WS has studied and conducted by many techniques, such as Maximal Matching [6],

which analyzed all possible word boundaries and select the best result to covered the ambiguous-word segmentation problem. Until now, the most popular Thai DB-WS program named Lexto [7] has been developed and distributed by the National Electronics and Computer Technology Center (NECTEC). MLB-WS is a non-dictionary word segmentation algorithm which required the training dataset that will be analyzed and extracted for some features used to defined the boundaries of words. The examples of Thai MLB-WS studies are [8–10] etc. Haruechaiyasak [11] compared the efficiency of DB-WS with 4 techniques of MLB-WS. The result showed that the DB-WS provided the highest accuracy in word segmentation.

2.2 Ranking Trie

Trie is a tree-like data structure created from words in a dictionary to serve the word's searching purpose. Each node of Trie is a character located as a child node of the previous character from the first character to the last character of each word. Traditional Trie may be created by any sequence of words or sorted by alphabetical order. Figure 1 (a) showed a traditional Trie of alphabetical order words: ”able”, ”always”, ”and”, ”angle”, ”angry”. Although Trie can reduce the size of a dictionary for searching, the arrangement of words of Traditional Trie in this manner is not effective, because all words can be found by the average probability. To improve the efficiency of word's searching on Trie, we propose the Ranking Trie [12] that rearrange all words by descending order of Word Usage Frequency (WUF). This technique provides the more frequently used words located at a higher level and could be found earlier than the fewer used words. Figure 1 (b) showed a Ranking Trie of words that are ordered by word usage frequency [13] as: ”and”, ”always”, ”able”, ”angry”, ”angle”.

2.3 Soundex

Soundex, a phonetic coding algorithm for names, was proposed by Robert C. Russell and Margaret K. Odell in 1918 [14]. The aim of using Soundex is to allow searching for the names which have the same Soundex even if they have different spelling. Currently, Soundex is wildly used in many languages. In the case of Thai, various encoding algorithm of Soundex were presented such as Lorchirachoonkul [15], Udompanich [16], and Wongtaweesap [17].

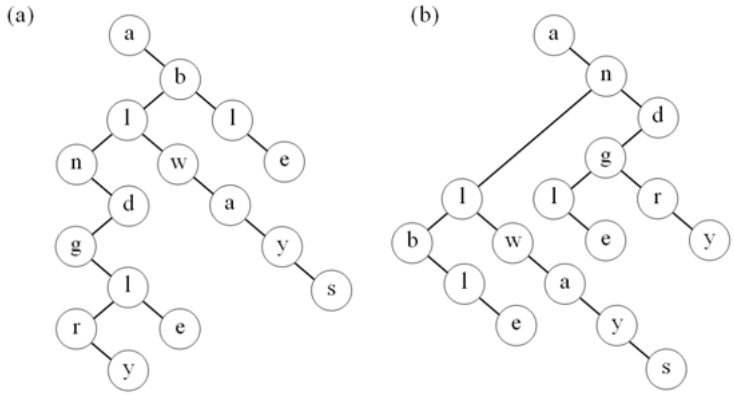


Fig. 1: (a) Traditional Trie structure and (b) Ranking Trie structure

2.4 Completed Soundex

Completed Soundex is a soundex which encoding all word's components including initial consonant, vowel, final consonant, tone, and cluster characters to provide a completed phonetic code that can be applied for misspelling correction by similarity analysis of words. The Completed Soundex code consists of one or more components called "Syllable code" that has 7 digits, divided into 5 subcomponents including initial consonant code, vowel code, final consonant code, tonal code, and cluster characters code as shown in Figure 2.

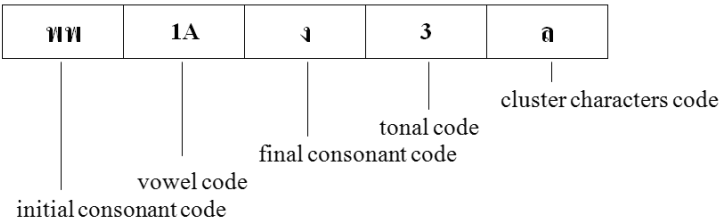


Fig. 2: Components of the Completed Soundex code

The number of syllable codes equal to the number of syllables of the word that is encoding. Coding values of the initial consonant, vowel, final consonant, tone, and cluster characters are showed as Table 1 to Table 3.

Table 1: Soundex coding values of the initial consonants.

Initial consonants	Soundex code	Initial consonants	Soundex code	Initial consonants	Soundex code
ก	กก	ด, ฎ	ดด	ฟ	ฟฟ
ค, ก, ฆ	คค	ต, ฏ	ตต	ฝ	ฟฝ
ข, ฃ	คข	น, ณ	นน	ม	มม
ง	งง	ท, ธ, ฑ, ฒ	ทท	ย, ญ	ยย
จ	จจ	ร, ฤ	ทร	ว, ฬ, ฬ, ฦ, ฦ	ววว
ช, ฌ	ชช	บ	บบ	ว	ววว
ฉ	ชฉ	ป	ปป	ฮ	ฮฮ
ซ, ทร	ชช	พ, ภ	พพ	ห	ฮห
ศ, ษ, ส	ชศ	ผ	พผ	อ	ออ

Table 2: Soundex coding values of the vowels.

Vowel	Vowel Soundex code	Additional Initial consonant	Additional final consonant	Vowel	Vowel Soundex code	Additional Initial consonant	Additional final consonant
◌ะ, ◌ั	1A			เ◌อะ	9Q		
◌า	1B			เ◌อ, เ◌ิ	9R		
◌ิ	2C			เ◌ียะ, เ◌ีะ	AS		
◌ี	2D			เ◌ีย, เ◌ี	AT		
◌ึ	3E			เ◌ือะ, เ◌ือะ	BU		
◌ื, ◌ือ	3F			เ◌ือ, เ◌ื	BV		
◌ุ	4G			◌ัวะ	CW		
◌ู	4H			◌ัว, ว	CX		
เ◌ะ, เ◌็	5I			ฤ	3E	ร	
เ	5J			ฤๅ	3F	ร	
แ◌ะ, แ◌็	6K			ฦ	3E	ล	
แ	6L			ฦๅ	3F	ล	
โ◌ะ	7M			◌ำ	1A		ม
โ	7N			ไ, ใ	1A		ย
เ◌าะ, ◌็	8O			เ◌า	1A		ว
อ	8P			รร	1A		น

Table 3: Soundex coding values of the tonal marks and cluster characters.

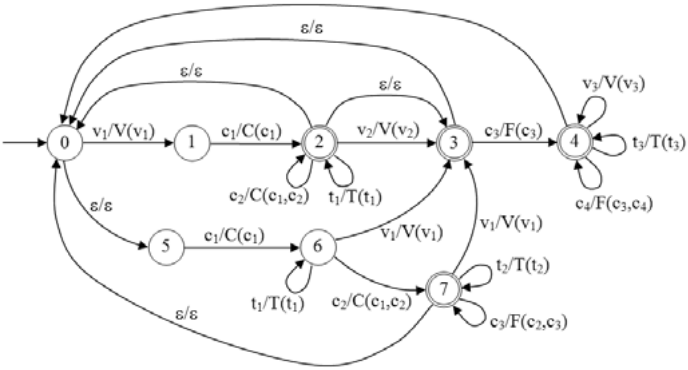
Tonal marks		◌่	◌้	◌๊	◌๋
Soundex code	0	1	2	3	4

Cluster charater	ร	ล	ว
Soundex code	ร	ล	ว

Table 4: Soundex coding values of the final consonants.

Final consonants	Soundex code
ก, ข, ฃ, ค, ฅ, ฆ, จ, ช, ซ, ฌ, กร, คร	ก
ค, จ, ช, ซ, ฎ, ฏ, ฐ, ฑ, ฒ, ด, ถ, ท, ธ, ศ, ษ, ส, ฉ, ฌ, ตร, ทร	ค
บ, ป, ภ, พ, ฟ, ปร	บ
ง	ง
น, ญ, ณ, ร, ล, ฬ	น
ม, ำ	ม
ย, ไ, ใ	ย
ว, เ-า	ว

The encoding process of Complete Soundex can be shown as a Nondeterministic Finite Automaton with output in Figure 3. The process starts from state 0, then move to next proper states according to the sequence of each component in the syllable, and produce output by the specific encoding function such as C(), V(), and F(). Some examples of state transition, encoding outputs of Complete Soundex encoding are shown in Table 5.



c1, c2, c3, c4	is an initial consonant
v0, v1, v2, v3	is a vowel
t1, t2, t3	is a tonal mark
C()	is an initial consonant and cluster characters encoding function.
V()	is a Vowel encoding function.
F()	is a final consonant encoding function.
T()	is a tonal encoding function.

Fig. 3: Nondeterministic Finite Automaton of Completed Soundex encoding process

Table 5: Examples of state transition and Complete Soundex encoding output of words.

Word	Phonetic	State transition	Output	Soundex
ก้าน	kâ:n	01223	C(ก)T(้)V(า)F(น)	ก11Lน2-
อัศรา	ʔat̄ sa:	012333	C(อ)V(ั)F(ศ)F(ศร)C(ศ)V(า)	อ11Sศ0-ศ11L-0-
มารค	ma:k̄	01233	C(ม)V(า)F(ร)F(รค)	ม11L-ก0-
สวรรค์	sawăn	01233	C(ส)C(สว)F(รร)F(รร)	ซ21S-0-ร11Sน0-
กลับ	klàp̄	01223	C(ก)C(กล)V(ั)F(บ)	ก11Sบ0ล
สามารถ	sǎ: mâ:t̄	0123333	C(ส)V(า)F(ม)C(ม)V(า)F(ร)F(รถ)	ซ21Lม0-ม11Lศ0-
กมล	ka mo:n	0123	C(ก)C(กม)F(มล)	ก11S-0-ม11Sน0-
เครื่อง	kʰuâŋ	00122233	V(เ)C(ค)C(คร)V(ื)T(่)F(อ)F(ง)	ค1BLง1ร
แกน	kɛ:n	0012	V(แ)C(ก)C(กน)	ก16Lน0-
อำนาจ	ʔam nâ:t̄	0120123	C(อ)V(ำ)C(น)V(า)F(จ)	อ11Sม0-น11Lศ0-
ทราย	sa:j	01223	C(ท)C(ทร)V(า)F(ย)	ซ11Lย0-

2.5 Completed Soundex Similarity Values

The similarity value between two words S(X, Y) can be defined by the Jaccard's coefficient which can be summarized as the following function.

$$S(X,Y) = \begin{cases} \sum_{i=1}^{n} \frac{sim(x_i, y_i)}{n} \\ sim(x_i, y_i) = \begin{cases} 1 \; if \; \exists \; x_i, = y_i \\ 0 \; if \; \exists \; x_i \neq y_i \end{cases} \\ 1 \leq i \leq n \\ n \; = \; max(x_{len}, y_{len}) \end{cases}$$

X, Y	is the words that are compared
x_i, y_i	is the ith Soundex code digit of X and Y respectively
$sim(x_i, y_i)$	is the similarity between the ith Soundex code digit of X and Y.
x_{len}, y_{len}	is the length of a Soundex code of X and Y respectively.
n	is the maximum length of a Soundex code of X and Y.

The example of similarity in each Soundex code digit between two words, and the similarity value calculation shown in Figure 4.

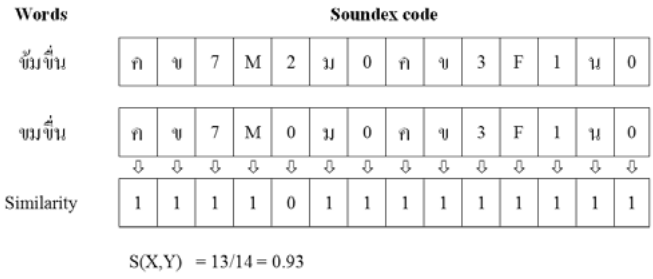


Fig. 4: Complete Soundex similarity value

3 Research Concept

This research presents a new algorithm that improved the efficiency of word segmentation and corrects the misspelling words which are the major problems in Natural language processing. As shown in Figure 5, the researchers developed a model named "Thai Language Segmentation by Automatic Ranking Trie with Misspelling Correction (TLS-ART-MC)" that applied two new techniques: Ranking Trie and Completed Soundex. The data used in these research are 6,579 text files collected from various sources of data in both online and offline covered 10 content's types, including economy, society, politics, health, education, agriculture, entertainment, sports, technology and IT, and others. The data was divided into 2 sets: 1) Learning set which are 5,879 text files randomly selected and used for word segmentation model developing, 2) Test set, the rest 700 text files are used for testing and performance evaluation.

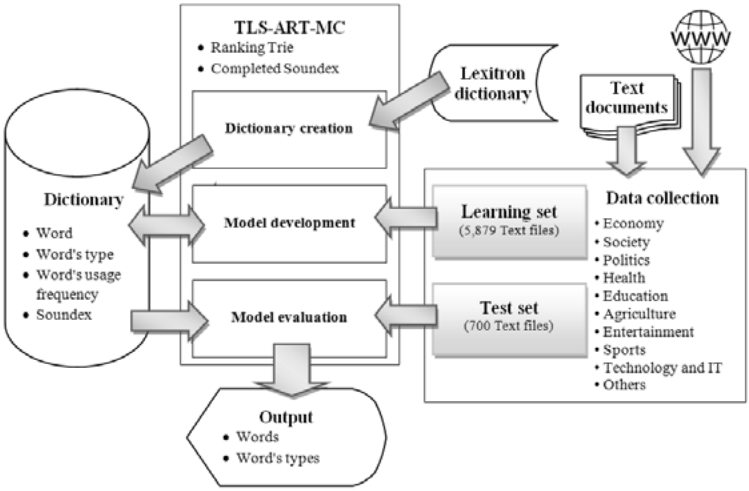


Fig. 5: Conceptual framework of the research

3.1 Dictionary Creation

As shown in Figure 6, the researchers used the Lexitron dictionary [18] which consists of 42,222 words to create the model's dictionary. Each word is inputted to create it's Completed Soundex and saved into the dictionary by Dictionary update process.

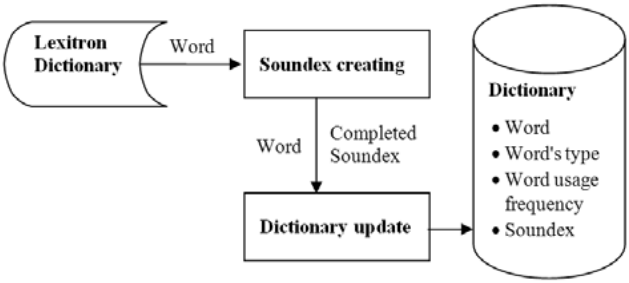


Fig. 6: The process of creating a dictionary

3.2 TLS-ART-MC Model

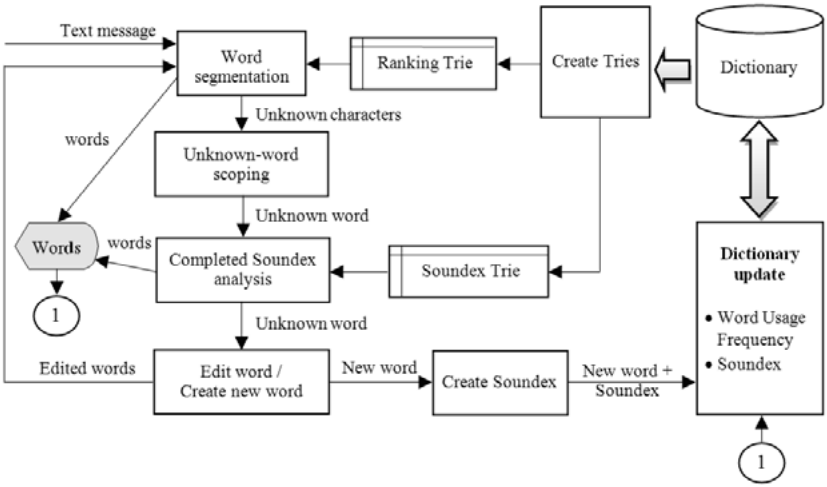


Fig. 7: The TLS-ART-MC model

As shown in figure 7, text messages from the learning set are inputted to the word segmentation module to parsed with the Ranking Trie created from data in the dictionary then output the words as a result. In the case of unknown characters are found, these unknown characters were passed to the Unknown-word scoping module to analyze and define the beginning and end of the possible unknown word by some predefined scoping rules that according to the Thai syllable spelling principle. Then the unknown word is parsed with the

Soundex Trie to define the most similar word which will be used to replace the unknown word. In case of no words that have similarity values higher than 0.8, the unknown word will be passed to the Edit word/Create new word module to generate some dialogues that help user edit the misspelling word, and send back to Word segmentation module, or add new words and create Completed Soundex codes for dictionary update.

4 Model Development

In this step, the researchers created all components of the model including the dictionary and modules Then input the learning set which are 5,879 text files for testing and use the results to improve the model.

5 Experiment and Performance Evaluation

The experiment in this research was conducted for the performance testing on 2 major parts of the model that are word segmentation and misspelling correction. For word segmentation, the researchers input the test set, which is a total of 700 text files, into the model for word segmentation and then compares the results with the word segmentation using the LexTo program and evaluates the efficiency with the precision, recall, and F-measure values. For the misspelling correction by Completed Soundex, The researchers input a total of 1,014 misspelling words cover 6 types of typo errors including Excess of alphabets, Missing of alphabets, Repetition of alphabets, Typo error, Misplacement of alphabets, Slang words and Mixed type error [19], into the completed Soundex analysis module to encode for Soundex and parsed to the Soundex Trie which created from words in the dictionary to find the most similar words that can be used to replace the misspelling words. The efficiency values are calculated as precision, recall, and F-measure values.

6 Experimental Results

6.1 Performance of Word Segmentation

The results of the comparative experiment between word segmentation by TLS-ART-MC and LexTo including number of True Positive (the words with correct segmentation), False Positive (the wrong words that are wrong segmented and presented as correct words), False Negative (the correct words that are wrong

segmented and presented as wrong words), and True Negative (the unknown words that are correctly identified) are shown in Table 6.

Table 6: Word segmentation results compared between TSL-ART-MC and Lexto

Segmentation algorithms	True Positive	False Positive	False Negative	True Negative
TLS-ART-MC	190,675	8,146	4,408	704
LexTo	189,758	8,414	4,516	754

The Accuracy, Precision, Recall, and F-measure values are calculated and shown in Table 7.

Table 7: Performance evaluation values of TSL-ART-MC and Lexto

Segmentation algorithms	Accuracy	Precision	Recall	F-measure
TLS-ART-MC	0.938	0.959	0.977	0.968
LexTo	0.936	0.958	0.977	0.967

6.2 Performance of Correcting Misspelled Words

The results of correcting misspelled words by the Complete Soundex are shown in Table 8.

Table 8: The results of correcting misspelled words by the Complete Soundex code

Typo error's type	Number of words			
	Tested words	True Positive	False Positive	False Negative
Excess of alphabets	178	159	23	19
Missing of alphabets	180	164	18	16
Repetition of alphabets	179	174	8	5
Typo error	183	172	15	11
Misplacement of alphabets	182	175	10	7
Slang and Mixed type error	112	107	6	5
Total	1,014	951	80	63

The Precision, Recall, and F-measure calculated from the results in Table 8 are 0.92, 0.94 and 0.93 respectively.

7 Conclusion

According to the results in the previous section, although the performance of word segmentation by TLS-ART-MC with Ranking Trie algorithms can reduce the size of the dictionary and the number of parsing tasks effectively, there still have an interesting question: In case of using with other jobs, how many suitable and sufficient words should be stored in the dictionary? The researcher found that, when testing by big data, new words that are the basic-used vocabulary are slightly increased while compound words, abbreviations, specific names, foreign spelling terms, and slang are increased with higher rate and not suitable to be stored in the dictionary due to too much number of occurrences. For this reason, further research should be conducted on this issue in order to provide the word segmentation process that can define the boundaries of these words without saving in the dictionary.

References

[1] Thairatananond, Y.: Towards the Design of a Thai Text Syllable Analyzer, *Master Thesis. Asian Institute of Technology*. Bangkok, Thailand, 1981

[2] Chanyapornpong, S.: A Thai Syllable Separation Algorithm, *Master Thesis. Asian Institute of Technology*. Bangkok, Thailand, 1983

[3] Pooworawan, Y.: Dictionary-based Thai Syllable Separation, In: *Proceedings of the Ninth Electronics Engineering Conference*, 1986

[4] Raruenrom, S.: Word segmentation by dictionary, *Department of Computer Engineering*, Chulalongkorn University. Bangkok, Thailand, 1991

[5] Smith, P: *Applied Data Structures with C++*, Jones and Bartlett publisher. Massachusetts, USA, 2004

[6] Sornlertlamvanich, V.: Word segmentation for Thai in machine translation system, *National Electronics and Computer Technology Center*.Bangkok, Thailand, 1993

[7] National Electronics and Computer Technology Center. *Thai Lexeme Tokenizer : LexTo*. Retrieved July 10, 2019, from `http://www.sansarn.com/lexto/`

[8] Chaloenpomsawat, P.: Feature-Based Thai Word Segmentation, *Chulalongkorn University*. Bangkok, Thailand, 1998

[9] Kawtrakul, A., Thumkanon, C. and Seriburi S.: A Statistical Approach to Thai Word Filtering, In: *Proceedings of the 2nd Symposium on Natural Language Processing*, Bangkok : 398-406, 1997

[10] Suwannawach, P.: Thai Word Segmentation Improvement using Maximum Matching and Tri-gram Technique, *Master Thesis of King Monkut's Institute of Technology Ladkrabang*.Bangkok, Thailand, 2012

[11] Haruechaiyasak, C., Kongyoung, S. and Dailey, M.: A Comparative Study on Thai Word Segmentation Approaches, *IEEE Proceedings of ECTI-CON 2008. 5th International Conference*, 125â128, 2008

[12] Tapsai, C., Meesad, P., and Haruechaiyasak, C.: TLS-ART: Thai Language Segmentation by Automatic Ranking Trie, Paper presented at: *The 9th International Conference Autonomous Systems*. Cala Millor, Spain, 2016

[13] Corpus of Contemporary American English. *Word frequency data*. Retrieved July 19, 2019, from `https://www.wordfrequency.info/free.asp`

[14] Odell, K.M. and Russel, C.R.: *US Patents 1261167 (1918) and 1435663 (1922)*

[15] Lorchirachoonkul, V. and Kudinpan, C.: Research Report: Thai Soundex and Syllable Segmentation Algorithm, *Journal of Universal Computer Science*, National Institute of Development Administration. Bangkok, Thailand, 1980

[16] Udompanich, W.: String searching for Thai alphabet using Soundex compression technique, *Department of Computer Engineering, Graduate School, Chulalongkorn University.* Bangkok, Thailand, 1983

[17] Wongtaweesap, O.: An Improvement of Thai/English Transliterated Word Encoding for Cross-language Retrieval by Syllable Segmentation of Phonetic Codes, *Department of Computer Engineering, Graduate School, Chulalongkorn University.* Bangkok, Thailand, 2006

[18] National Electronics and Computer Technology Center. *Thai Lexeme Tokenizer: Lexitron dictionary*. Retrieved July 10, 2019, from `http://www.sansarn.com/lexto/download/lexitron.txt`

[19] Tapsai, C.: Analysis of Patterns and Causes of Misspelling and Slang Words for Natural Language Processing, In: *Proceedings of the 442nd International Conference on Science, Technology and Management*. Moscow, Russian Federation: 25-30, 2018

Fundamentals and Theory

An Information Flow Model to Support NIS Mandated Reporting

Gerald Quirchmayr, Veronika Kupfersberger,
Gregor Langner and Thomas Schaberreiter

University of Vienna, Faculty of Computer Science
Multimedia Information Systems Research Group
Währinger Straße 28, A-1090 Vienna, Austria

Abstract: After briefly motivating the research and listing the major challenges resulting from the NIS Directive, this contribution presents an information flow model which is, amongst other goals, aimed at supporting reporting obligations mandated by the NIS Directive. The model and its components are then described, followed by an example of the models information sharing and exchange component as it is currently used in the CS-AWARE project. The paper finishes with an outlook and conclusions.

1 Introduction

The European Union has introduced a series of new legislation [1, 2], aimed at countering the growing number and sophistication of threats against security and privacy. With the criminal and political landscapes changing for the worse [3], this consolidated European legislative response comes at the right moment, trying to balance the necessities of protecting ICT infrastructures with privacy requirements. Especially against the background of Industry 4.0 becoming a major cornerstone of the European economy [4], this new legislation can be expected to have a major societal impact. With critical infrastructures and significant digital services being the focus of the NIS Directive, the protection of vital societal services now receives the much needed legal attention [10].

2 Major New Obligations Introduced by the NIS Directive and Arising Challenges

The major obligations introduced by the NIS Directive are concerned with the protection of vital infrastructures and reporting duties in case of major incidents.

The primary goal of the legislation is to enhance the resilience of critical infrastructures and to establish an early warning mechanism that allows a coordinated response. As a reliable infrastructure and a dependable network between trusted partners are crucial for the successful implementation of truly European information sharing, protection and coordinated response mechanisms, a network of trusted nodes is needed. In this situation the obvious choice was made - giving national and sector CSIRTs a central role in coordinating these efforts. These elements of the Directive were translated into national legislation in European Union Member States over the past years and are now being applied. Especially the resulting mandatory reporting duty for significant cyber incidents is expected to be a major game changer regarding the situational awareness in Member States and ultimately across the whole European Union. One of the resulting major challenges for affected organizations now is to put in place a situational awareness tool that allows them to identify, detect and report indicators of a major attack and to correlate events across the organization, which is even more difficult in case of a distributed ICT environment.

3 An information Flow Model to Support NIS Mandated Reporting

Given the challenges described in the previous chapter, the need for a support tool is obvious. In order to build a sustainable approach, an information flow model [5] was developed to feed a situational awareness framework, including the reporting functionality mandated by the NIS Directive. In order to be effective, this information flow model is aimed at helping to identify attacks, supporting the application of counter measures and providing enough information about an attack to allow a meaningful reporting. The architecture is described in Figure 1.

As can be seen from the illustration above, the model comprises several relatively independent components. This approach was chosen to allow a flexible execution, because not every organization needs all of the components. The functionality of these components is as follows [6]: As basis for further analysis, a System Dependency Analysis is carried out as starting point. **The System Dependency Analysis** is performed by combining the Soft Systems Methodology and the GraphingWiki, resulting in a strategic implementation process for the concerned organization. Based on the pilot analyses, guidelines for future System Dependency Analyses will be developed. The next step is focused on **Data Collection**. The Data Collection component will be responsible for developing the data collectors for system specific data as well as external sources.

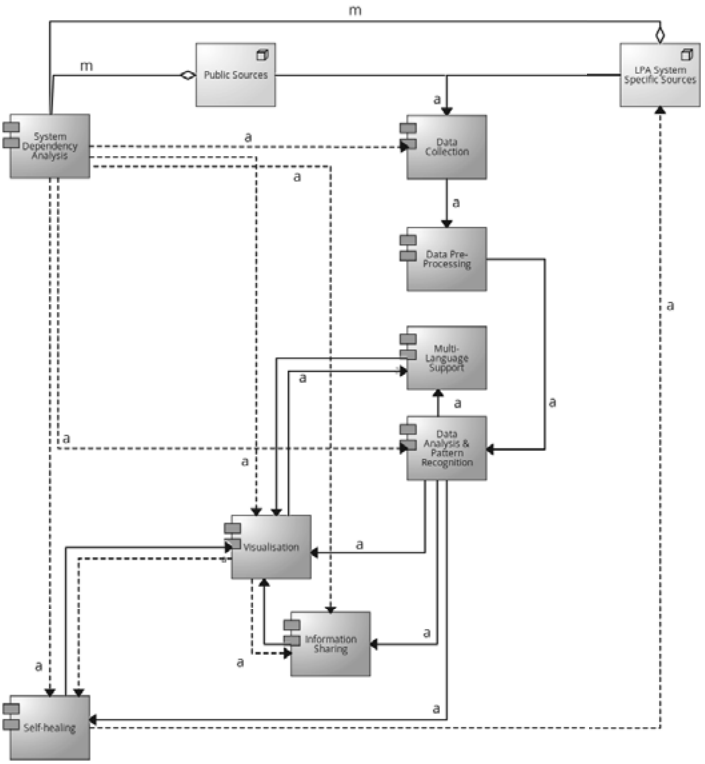


Fig. 1: The CS-AWARE Information Flow Model [5]

The Cybersecurity Information Exchange module is responsible for sharing information on detected attacks with authorities, according to the NIS regulations. **The Cybersecurity Information Exchange** will allow the user to individually authorize any transmission before it occurs. A **Visualization** component covers the final data manipulations required for graphically representing the collected information as well as the construction of the user interface. The **Self-Healing** component will receive information from the Data Analysis module and compose Security Rules based on the detected incident. These rules can then be applied to the systems by the respective IT departments. The **Data Pre-Processing** component executes pre-defined strategies, one of them being the **Natural Language Processing** for Information Extraction, which is a simplification process

of textual information. Other possibilities would include the simple filtering of known irrelevant data or the transformation of data formats. Due to the European context of CS-AWARE, the final UI should include not only easily understandable visualizations but all text in either the native language of the end user or English. **Multi-Language Support** is therefore required.

As security solutions do heavily depend on an organizations policies [7], any implementation of the model will have to cover on site, cloud based and hybrid variations. That is the major reason why Docker containers were chosen as basis for the system architecture. Another advantage of this approach is a high flexibility in orchestrating execution paths that usually vary from organization to organization.

4 Application of the Information Flow Model in Local Public Administrations [8]

The first context in which this information flow model and the related framework were applied is the CS-AWARE project, which is aimed at providing local public administrations with a tool that creates the necessary awareness in case a cyberattack occurs. As especially larger local public administrations also run critical infrastructure services that come under the NIS Directive, such a framework and the related tool support are a highly welcome resource. The major user group to be provided with awareness raising information are system and information security administrators, because they are the ones who most need a "big picture" of an attack situation to be able to identify the target of an attack and the modus operandi of an attacker. As important as a forensic evidence collection is for attributing an attack to an attacker, the major goal of this information flow model is to support those who are charged with defending an attacked system. The primary use cases therefore are the detection of vulnerabilities and the identification and classification of an attack. Both goals are in line with the NIS Directive, strengthening existing defenses and improving early warning capabilities. In this context the visualization component acquires a central role, pointing the user directly to where a problem occurs. Regarding the incident reporting obligation introduced by the NIS Directive, the **Cybersecurity Information Exchange** module plays a core role. The docker container architecture used in the project is shown in Figure 2.

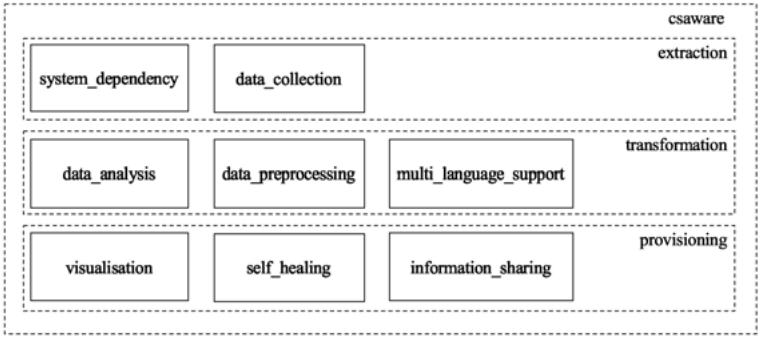


Fig. 2: CS-AWARE Docker Container Architecture [6]

Table 1: Docker Compose Commands and their Parameters used in CS-AWARE [6]

Commands	Parameters	Definition
build	context dockerfile	Any information relevant at build-time is specified here Location of built container is specified. If required, the path to an individual dockerfile for this specific container can be specified here
restart		Containers can be told to automatically restart after shutdown either on failure, always or unless-stopped
networks		Either the default network created by Docker Compose is used or individual networks can be defined
	aliases internal	Container can have aliases in each network they are allocated to In internal networks, only other containers of the same network can gain access
logging ports		Activates the automated logging function of Docker Individual ports can be defined for each container, which are opened and can be accesses by other containers or external software
image		Can either be a new name for the created container or an existing image can be loaded here

Table 1 gives an overview of the docker compose commands and their parameters as used in the CS-AWARE project.

In order to be open for future developments, the STIX/TAXII [9] standards are applied as basis for communication and information sharing, both internally and externally. As these two standards now also starting to be introduced in smart manufacturing infrastructures, embedding the developed information flow model in cyber physical systems becomes a viable option.

5 Outlook and Conclusion

Given the reporting obligations introduced by the NIS Directive, tool-supported models such as the one presented in this paper will become a common necessity for critical (information) infrastructure protection. Defensive measures, especially the fast identification and detection of threats, will play a decisive role in making our ICT infrastructures more resilient and will be essential for establishing an EU wide early warning and response coordination system [10]. Especially the reporting network which is building on national and sector CERTs is expected to become a major game changer once it is fully operational [11, 12]. This is another reason why the presented information flow model has a high potential beyond the public sector.

6 Acknowledgements

The authors would like to thank the EU H2020 project CS-AWARE (grant number 740723) for supporting the research presented in this paper.

References

[1] The Directive on security of network and information systems (NIS Directive) (2016b) https://ec.europa.eu/digital-single-market/en/network-and-information-security-nis-directive

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). (2016a)

[3] The 2018 Internet Organised Crime Threat Assessment, Europol, 2018, available on https://www.europol.europa.eu/sites/default/files/../iocta2018.pdf

[4] Bogoviz A.V. (2019) Industry 4.0 as a New Vector of Growth and Development of Knowledge Economy. In: Popkova E., Ragulina Y., Bogoviz A. (eds) Industry 4.0: Industrial Revolution of the 21st Century. Studies in Systems, Decision and Control, vol 169. Springer, Cham, ISBN 978-3-319-94309-1, https://doi.org/10.1007/978-3-319-94310-7_8

[5] Kupfersberger, V., Schaberreiter, T., Quirchmayr, G., Security-Driven Information Flow Modelling for Component Integration in Complex Environments, *in Proceedings of the 10th International Conference on Advances in Information Technology* (2018), ISBN 978-1-4503-6568-0.

[6] Kupfersberger, V., The CS-AWARE Information Flow Model, Masterâs Thesis at the Univeristy of Vienna, 2018, supervised by Gerald Quirchmayr and Thomas Schaberreiter.

[7] Safa, S.N.,Von Solms, R., Furnell, S., Information security policy compliance model in organizations, in Computers & Security, Volume 56, February 2016, Pages 70–82, https://doi.org/10.1016/j.cose.2015.10.006

[8] EU H2020 project CS-AWARE (grant number 740723), https://cs-aware.eu/

[9] Cybersecurity and Infrastructure Security Agency, Information Sharing Specifications for Cybersecurity, https://www.us-cert.gov/Information-Sharing-Specifications-Cybersecurity

[10] European Union Cyber Security Strategy, https://ec.europa.eu/digital-single-market/en/cyber-security

[11] ENISA on CSIRTs network, https://www.enisa.europa.eu/topics/csirts-in-europe/csirts-network

[12] ENISA Study on CSIRT landscape and IR capabilities in Europe 2025, V 1.0 — February 2019, https://www.enisa.europa.eu/publications/study-on-csirt-landscape-and-ir-capabilities-in-europe-2025

Automatic Evaluation of Programming Tasks for Blended Learning and E-Assessment

Brigitte Hass[1,2], Chunrong Yuan[1] and Zhong Li[2]

[1]Autonomous Systems Lab
Faculty for Information, Media and Electrical Engineering
TH Köln – University of Applied Sciences, Cologne, Germany

[2]Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

With the fast development of computer and internet technology, blended learning has become an efficient methodology used pervasively in both academic teaching and general purpose training courses, where interactive frontal teaching and computer-supported self-learning in multimedia environments are integrated seamlessly. From the perspectives of learners as well as teachers, convenient ways for performance checking are needed. On the one side, students can make self-assessments so as to adjust their learning focus and make improvements based on feedbacks. On the other side, teachers can design contents, material and activities tailored specifically to the need of students. This raises the question of E-assessment, i. e., computer-based automatic evaluation of learning outcomes.

In the field of science and engineering education, it is necessary for the students to learn, design and develop algorithms and computer programs. Tasks in such courses are open questions and there are usually more than one or even an infinite number of possible solutions. Here the major challenge of E-Assessment lies in the automatic checking of program codes written by students. Such an E-Assessment system should decide among others whether the codes are programmed properly, whether there are mistakes which could lead to compiler or run-time errors, whether the programs work as expected and are able to pass various test conditions, etc.

Although there exists already research work dealing with this problem, by using either static or dynamic analysis methods, there are various restrictions and limitations. Regarding static code analysis, only simple programs with very short length as well as simple logic and structures can be evaluated, as checking is based mostly on textural information, e. g., by line-for-line comparison between

the solution of a student and a standard program code. For dynamic or runtime error checking, the situation is even worse. So far, only elementary functions have been realized as research prototypes, whose performances are either unsatisfactory or not studied at all.

The focus of this work lies in the automatic assessment of learning outcome in programming techniques. Based on the E-learning platform ILIAS[1] which uses a client-server architecture, an E-Assessment system has been developed which is capable of checking programming tasks in Java. This E-Assessment system is able to perform both static and dynamic code analysis. Here, ILIAS is mainly used for the realization of front-end functions such as material uploading, task assignment, code submission and feedback delivery. The underlying assessment functions run in the background. Some of the functions are: Static code analysis based on Checkstyle[2], dynamic code analysis based on JUnit and JavaRE test, and plagiarism detection based on JPlag[3]. Furthermore, the system is able to give statistical analysis of the performances of all students in a course, not only for individual programming tasks, but also for the whole assignments within particular learning periods. As a consequence, the E-Assessment system can be used not only for the purpose of exercises, but also for periodical examination situations including final exams.

Our ongoing work aims at achieving integrated code checking which combines the static and dynamic code analysis processes. Another future work is the extension of our current E-Assessment system for intelligent tutoring purposes, where students will receive powerful feedbacks, get step-by-step guidances, and perform multi-modal interactions and collaborations in both formal and informal learning environments. Such kinds of intelligent and autonomous systems are achievable with further development and effective use of AI (Artificial Intelligence) and AR (Augmented Reality) technologies.

[1] `https://ilias.th-koeln.de`
[2] `https://checkstyle.sourceforge.io`
[3] `https://jplag.ipd.kit.edu`

Matrix-based Feature Selection for Dynamic Multi-source Hybrid Data

Yanyong Huang[1,2,3]

[1]Chair of Communication Networks, FernUniversität in Hagen, Germany

[2]School of Information Science and Technology,
Southwest Jiaotong University, Chengdu, China

[3]School of Statistics,
Southwestern University of Finance and Economics, Chengdu, China

Abstract: In real-life applications, the collected data always are characterized by multiple information sources and different types of features including boolean, categorical and numerical. In this paper, we call this type of data as multi-source hybrid data. Feature selection can reduce the redundant and irrelevant information, which is very important for multi-source data classification task. Existing feature selection methods based on rough sets focus on single source data. They could not be directly applied to dynamic multi-source hybrid data with the variation of data sources. To address this issue, we proposed a novel information entropy based on multi-kernel fuzzy rough set, which can efficiently characterize the uncertainty measure for multi-source hybrid data. Then, the information entropy is constructed by multi-source composite relation matrix and different matrix operators. Given this effective classification capability, the incremental feature selection mechanisms and algorithms are developed for dynamic multi-source data under the addition and deletion of data sources. Finally, extensive experiments are carried out to verify the effectiveness of the proposed methods when comparing with other related feature selection algorithms. Experimental results show that the proposed feature selection method outperforms the related approaches in terms of classification accuracy and computing overhead.

SafeBigNum – Arbitrary-precision Arithmetic with Integrity Checking

Stefan Widmann

Abstract: Important applications like asymmetric cryptography require processing numbers exceeding the native bit width of arithmetic registers of the machines they are being run on. These numbers are called BigNum, BigInt or arbitrary-precision numbers and are being supported by a number of libraries. The demand for data confidentiality and therefore data encryption in safety-related applications is rising, leading to the requirement of checking integrity and timeliness of BigNums, which current libraries do not provide. Using the features of tagged memory architectures, a method for verification of integrity and timeliness of arbitrary-precision numbers is introduced.

1 Introduction

Arbitrary-precision arithmetics, also called BigNum or BigInt arithmetic, is being used in various applications where the operands exceed the processor's native arithmetic register width. One of the most important of these applications is asymmetric cryptography. There are several libraries available to provide developers with support for arbitrary-precision numbers. Since confidentiality of data is getting more and more important in safety-related applications, where systems are responsible for human lives, nature and investments, it is time to take a look at BigNums from a safety point of view. BigNum data structures in typical libraries consist of several management variables such as the number of words allocated for the number and the number of words being currently in use and an array or an pointer to an array holding the BigNum's actual data values. The most commonly used libraries do not implement any measures to be able to check integrity and timeliness of BigNums. In this paper an easy method for hardware-based integrity and timeliness checking is introduced using features of tagged memory architectures. Furthermore, an speed-optimized method for updating the integrity- and timeliness-checking tag information is provided.

2 Typical Data Structure Layout for BigNums

To illustrate the data structures typically being used to represent BigNums in memory, the "mp_int" data structure of the library LibTomMath of the LibTomCrypt project [9] shall be explained here. Listing 1 shows the C language datatype definition of mp_int based on [9].

Listing 1: mp_int data structure

```
typedef struct {
  int used, alloc, sign;
  mp_digit *dp;
} mp_int;
```

The data structure's different members are:

- *alloc*, which specifies the number of bytes allocated for the digits of the mp_int variable,
- *used*, denoting how many of the allocated digits are currently in use,
- *sign*, specifying whether the mp_int number is positive or negative and
- *dp*, being a pointer to the allocated digit memory words of type *mp_digit*.

The bit width of a single mp_digit can vary between 8 and 128 bits. A comprehensive list of different available BigNum libraries can be found at [13].

Figure 1 shows the exemplary memory layout of the positive BigNum 0x123456789ABCDEF08154711 with 32 bit wide digits.

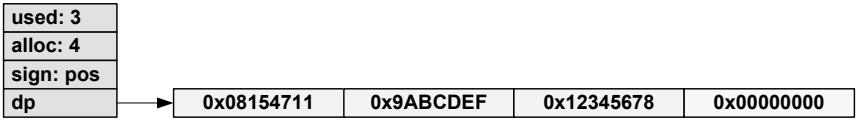


Fig. 1: Exemplary layout of a BigNum in memory

Since the width of BigNums usually exceed the bit width of the arithmetic registers of the processing machine, the BigNum library software must implement different calculation schemes for providing typical arithmetic operations on BigNum variables. The addition is shown in figure 2.

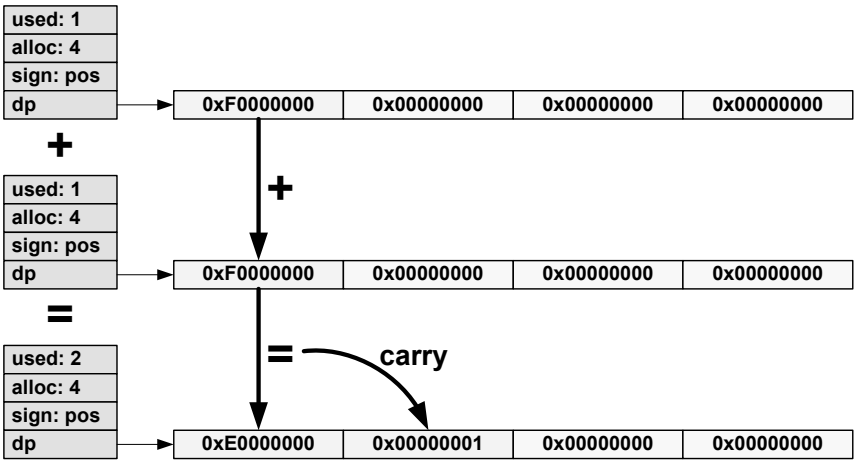


Fig. 2: Addition scheme for BigNums

It is easy to see in the data structure illustrated above, that there are no fields to keep any integrity or timeliness checking data.

3 BigNum-relevant Errors in Data Processing

Schiffel [8] extended Forin's error model [5] to cover the following error types relevant for this paper:

- corruption of data values and
- lost updates resulting in outdated operands.

In the special case of BigNums, corruption and missing timeliness does not only affect whole BigNum variables, but can affect single words within the number's memory structure, too.

Such errors must be detected and subsequently handled appropriately before they can lead to dangerous outputs that could endanger human lives, the environment or investments.

4 State of the Art

As state of the art, the hardware-based ECC- and parity integrity checking methods and the software-based ANBD coding are relevant to this paper.

4.1 Hardware-based Integrity Checking

Many x86-based servers use Extended-(72,64)-Hamming-Codes [6] to detect double bit errors and correct single bit errors. Another example for hardware-based integrity checking is the 1-bit parity of various STM32 microcontrollers such as the STM32F0 [10], which is used to detect corrupted data memory.

4.2 ANBD Coding

ANBD coding is a software-based arithmetic coding technique, first introduced as AN coding in [3] and extended to ANBD coding in [5]. It can be used to improve error detection on conventional architectures like x86 and ARM. Data values N are coded by adding an integrity check A, an identifier B and a time-stamp D resulting in the coded value x_c:

$$x_c = A \cdot x + B + D$$

The integrity of x_c can be checked using a given B and D by verifying, that the equation

$$x_c = A \cdot x + B + D \equiv B + D \mod A$$

is satisfied.

Using ANBD coding, corrupted or outdated operands can be identified as such. Some coded operations require complex corrections of their results, and some operations like divisions are problematic.

4.3 Drawbacks of the State of the Art

The ECC- and parity-based integrity checking schemes are not suitable for checking the integrity of data structures consisting of several memory words, since they only verify the integrity of a single memory word and do not provide any means for detecting outdated BigNums or parts thereof. ANBD coding could be implemented to check integrity and timeliness of BigNum variables, but the calculation effort for handling the coded numbers would be extremely high and operations like divisions would be hard to implement.

5 Safe Arbitrary-precision Numbers

Hardware-based integrity checking has been introduced long time ago in tagged memory architecture machines such as the AEG TR 4 [1]. These architectures add data value properties to memory words in form of tags. In the memory words, the tags are linked indivisibly to the data value and the words are stored, processed and transmitted as an atomic unit. Typical tag types are e.g. data type identifiers and per-word integrity checking data. Descriptor architectures like the Burroughs B5000 [7] added support for safe arrays by introducing array descriptor data types and hardware-supervised array element accesses. The safety and security benefits of using tagged memory architectures seem to have been forgotten for a long time, but have been recently used and extended in different architectures like for example lowRISC [2], SAFE [4] and ISMA [11]. The Data Specification Architecture DSA [12] introduced Timestep tags, carrying information about a data word's discrete update generation, allowing verification of the timeliness of operands and detection of corresponding errors like lost updates without the drawbacks of ANBD coding.

The Timestep tag data of the DSA [12] for verifying timeliness is focused on single memory words only and must be extended to support verification of multi-word data structures.

The new hardware-based data type "SafeBigNum" is introduced consisting of a SafeBigNum descriptor holding the information taken from the "MP_int" data structure shown before and several subsequent memory words holding the SafeBigNum's actual digits. Each memory word contains an integrity checking tag, allowing the verification of the word's integrity. Additionally, the descriptor and all array element memory words contain a Timestep tag, specifying the discrete update generation of the specific memory word.

An example of the new data type SafeBigNum is shown in figure 3, representing the number 0x123456789ABCDEF08154711. The descriptor specifies that 4 array elements are following in memory, of which 3 are being currently in use. In the shown example, each array element keeps 32 bits of the SafeBigNum's value. The sign of the number is positive. The Timesteps of the descriptor and all used array elements is 3. The integrity checking tag is not depicted in order to keep the figure simple.

To illustrate the handling of the new SafeBigNum data type and checking the integrity of the variables, figure 4 shows the addition of two SafeBigNums and

SafeBigNum descriptor	Array element 0	Array element 1	Array element 2	Array element 3
used: 3; alloc: 4; sign: pos	0x08154711	0x9ABCDEF	0x12345678	---
Timestep TS: 3	Timestep TS: 3	Timestep TS: 3	Timestep TS: 3	Timestep TS: -

Fig. 3: New data type SafeBigNum

the update of the Timestep tags. Although only array element 0 is changed during the addition, all Timestep tag contents are incremented.

SafeBigNum descriptor	Array element 0	Array element 1	Array element 2	Array element 3
used: 3; alloc: 4; sign: pos	0x08154711	0x9ABCDEF	0x12345678	---
Timestep TS: 3	Timestep TS: 3	Timestep TS: 3	Timestep TS: 3	Timestep TS: -
		+		
used: 1; alloc: 4; sign: pos	0x00000001	---	---	---
Timestep TS: 0	Timestep TS: 0	Timestep TS: -	Timestep TS: -	Timestep TS: -
		=		
used: 3; alloc: 4; sign: pos	0x08154712	0x9ABCDEF	0x12345678	---
Timestep TS: 4	Timestep TS: 4	Timestep TS: 4	Timestep TS: 4	Timestep TS: -
Integrity checking:	4 = 4	4 = 4	4 = 4	

Fig. 4: Exemplary addition of two SafeBigNums without error

While the integrity of the single memory words is checked by the hardware using the integrity check information not being shown in the figure, the overall integrity regarding timeliness and temporal coherence of the SafeBigNum can be verified by comparing the Timestep tag contents of all used array elements to the Timestep tag content of the descriptor as shown in listing 2.

Listing 2: SafeBigNum integrity check

```
FOR 0 ≤ i < used REPEAT
  IF Timestep(element[i]) ≠ Timestep(descriptor) THEN
    Generate_Exception;
  ENDIF
ENDFOR
```

The detection of an error that leads to an inconsistent SafeBigNum variable due to a lost update of an array element is shown in figure 5. Due to some kind of failure, updating the array element 0 fails. Using the integrity checking mechanisms for the single memory words, this error cannot be detected, since the per-word integrity of array element 0 is still given. Only the comparison of all Timestep tag contents reveals the outdated array element 0, giving the system the possibility to handle the detected error appropriately.

SafeBigNum descriptor	Array element 0	Array element 1	Array element 2	Array element 3
used: 3; alloc: 4; sign: pos	0x08154711	0x9ABCDEF	0x12345678	---
Timestep TS: 3	Timestep TS: 3	Timestep TS: 3	Timestep TS: 3	Timestep TS: -
		+		
used: 1; alloc: 4; sign: pos	0x00000001	---	---	---
Timestep TS: 0	Timestep TS: 0	Timestep TS: -	Timestep TS: -	Timestep TS: -
		= ERROR: array element 0 is not updated!		
used: 3; alloc: 4; sign: pos	0x08154711	0x9ABCDEF	0x12345678	---
Timestep TS: 4	Timestep TS: 3	Timestep TS: 4	Timestep TS: 4	Timestep TS: -
Integrity checking:	4 ≠ 3	Error is being detected!		

Fig. 5: Addition of two SafeBigNums with lost update of an array element

A lost update of the descriptor can be detected, too, as shown in figure 6, since the descriptor's Timestep does not match the Timestep of the first array element.

SafeBigNum descriptor	Array element 0	Array element 1	Array element 2	Array element 3
used: 3; alloc: 4; sign: pos	0x08154711	0x9ABCDEF	0x12345678	---
Timestep TS: 3	Timestep TS: 3	Timestep TS: 3	Timestep TS: 3	Timestep TS: -
		+		
used: 1; alloc: 4; sign: pos	0x00000001	---	---	---
Timestep TS: 0	Timestep TS: 0	Timestep TS: -	Timestep TS: -	Timestep TS: -
		= ERROR: descriptor is not updated!		
used: 3; alloc: 4; sign: pos	0x08154712	0x9ABCDEF	0x12345678	---
Timestep TS: 3	Timestep TS: 4	Timestep TS: 4	Timestep TS: 4	Timestep TS: -
Integrity checking:	3 ≠ 4	Error is being detected!		

Fig. 6: Addition of two SafeBigNums with lost update of the descriptor

The examples in figures 4 to 6 show, that even if only one digit of a SafeBigNum is updated during data processing, all the array elements have to be updated, too, in order to adjust the contents of their Timestep tags. Thinking of several-hundred-digit SafeBigNums, updating only one digit and subsequently updating all other digits just to increment their Timestep tags' contents introduces a significant overhead. This is why a speed-optimized method for updating Timestep data was developed.

6 Speed-optimized Updating of Timestep Tags

To reduce the described overhead of the naive Timestep tag update method described above, where all digits' Timestep tags have to be updated, a speed-optimized update method has been developed. Instead of updating all Timestep tags, only those of changed digits are updated during data processing. To

be able to verify the SafeBigNum's integrity, a checksum of the Timestep tags of all used digits is placed in the ∑Timestep tag ∑TS in the descriptor.

Possible checksum algorithms must fulfill the following requirements: they must be

- easy to implement in hardware and
- the checksum must be updatable without complete recalculation if only single memory words are being changed.

Suitable algorithms are e. g.

- a simple sum or
- XOR'ing

of the Timestep tag contents of the array elements.

The new speed-optimized SafeBigNum data structure is shown in figure 7. In the shown example, all Timesteps of the used digits of the SafeBigNum are added up and stored in the ∑Timestep tag of the descriptor. The ∑Timestep tag shall keep the sum of all elements' Timestep tags modulo 2^n, where n is the number of bits provided for storing the sum. This results in the ∑Timestep tag of the descriptor forming the residue class ring $\mathbb{Z}/2^n\mathbb{Z}$.

SafeBigNum descriptor	Array element 0	Array element 1	Array element 2	Array element 3
used: 3; alloc: 4; sign: pos	0x08154711	0x9ABCDEF	0x12345678	---
Timestep TS: 3; ∑TS: 9	TS: 3	TS: 3	TS: 3	TS: -

Fig. 7: Data structure of the speed-optimized SafeBigNum with checksum

The next example shown in figure 8 shall illustrate how the Timestep tags are updated. Only the first digit in array element 0 needs to be updated due to the addition and this element's Timestep tag is incremented from 3 to 4. The Timesteps of the other digits are not changed. The descriptor's Timestep is incremented, too, and the ∑Timestep tag is updated to the new expected Timestep sum of 10. During the integrity check of the SafeBigNum, all digits' Timesteps are added up and compared to the ∑Timestep tag of the descriptor. The method is illustrated by the pseudo code in listing 3.

SafeBigNum descriptor	Array element 0	Array element 1	Array element 2	Array element 3
used: 3; alloc: 4; sign: pos	0x08154711	0x9ABCDEF	0x12345678	---
Timestep TS: 3; ∑TS: 9	TS: 3	TS: 3	TS: 3	TS: -
+				
used: 1; alloc: 4; sign: pos	0x00000001	---	---	---
Timestep TS: 3; ∑TS: 3	TS: 3	TS: -	TS: -	TS: -
=				
used: 3; alloc: 4; sign: pos	0x08154712	0x9ABCDEF	0x12345678	---
Timestep TS: 4; ∑TS: 10	TS: 4	TS: 3	TS: 3	TS: -
Integrity checking:		∑TS = 4 + 3 + 3 = 10		

Fig. 8: Speed-optimized addition of two SafeBigNums

Listing 3: Speed-optimized SafeBigNum integrity check

```
Timestep_sum := 0;
FOR 0 ≤ i < used REPEAT
  Timestep_sum := Timestep_sum + Timestep(element[i])
ENDFOR
IF Timestep_sum ≠ ∑Timestep(descriptor) THEN
  Generate_Exception;
ENDIF
```

Again, the two error types ”lost update of array element” and ”lost update of descriptor” shall be detected when using the speed-optimized Timestep update method.

Figure 9 shows the addition with the update of array element 0 being lost. During integrity checking, the sum of the elements’ Timestep tag contents does not match the expected Timestep sum specified in the ∑Timestep tag of the descriptor. The error is being detected successfully and can be handled appropriately.

SafeBigNum descriptor	Array element 0	Array element 1	Array element 2	Array element 3
used: 3; alloc: 4; sign: pos	0x08154711	0x9ABCDEF	0x12345678	---
Timestep TS: 3; ∑TS: 9	TS: 3	TS: 3	TS: 3	TS: -
+				
used: 1; alloc: 4; sign: pos	0x00000001	---	---	---
Timestep TS: 3; ∑TS: 3	TS: 3	TS: -	TS: -	TS: -
=		ERROR: array element 0 is not updated!		
used: 3; alloc: 4; sign: pos	0x08154711	0x9ABCDEF	0x12345678	---
Timestep TS: 4; ∑TS: 10	TS: 3	TS: 3	TS: 3	TS: -
Integrity checking:		∑TS = 3 + 3 + 3 ≠ 10		

Fig. 9: Speed-optimized addition of two SafeBigNums with lost update of element

The detection of the second error type, "lost update of descriptor" is shown in figure 10. The sum of the Timestep tag contents of the used SafeBigNum array elements does not match the contents of the ∑Timestep tag of the descriptor. Again, the error is being detected successfully and can be handled appropriately.

SafeBigNum descriptor	Array element 0	Array element 1	Array element 2	Array element 3
used: 3; alloc: 4; sign: pos	0x08154711	0x9ABCDEF	0x12345678	---
Timestep TS: 3; ∑TS: 9	TS: 3	TS: 3	TS: 3	TS: -
	+			
used: 1; alloc: 4; sign: pos	0x00000001	---	---	---
Timestep TS: 3; ∑TS: 3	TS: 3	TS: -	TS: -	TS: -
	=	ERROR: descriptor is not updated!		
used: 3; alloc: 4; sign: pos	0x08154712	0x9ABCDEF	0x12345678	---
Timestep TS: 3; ∑TS: 9	TS: 4	TS: 3	TS: 3	TS: -
Integrity checking:		∑TS = 4 + 3 + 3 ≠ 9		

Fig. 10: Speed-optimized addition of two SafeBigNums with lost update of descriptor

7 A Proposal for Software-based SafeBigNums

The SafeBigNum structure and integrity checking methods shown and described in this paper are targeted on implementation in tagged memory architectures like SAFE [4], ISMA [11] and DSA [12]. They are not suitable for software-based implementations, since the tags and the data values will most likely not reside within single memory words and thus being updated sequentially. Due to this sequentiality, the inherent coherency of tags and data values, which tagged memory architectures are able to provide, is lost when being implemented in software and being run on conventional architectures.

However, if no tagged memory architecture is available, integrity checking of BigNums is still advisable. Figure 11 shows the proposal of an extension of existing BigNum data structures in BigNum libraries to SafeBigNums in order to detect corrupted and outdated operands.

The BigNum data structure is extended by two fields in order to provide the same error detection capability of the tagged memory architectures' solution:

- a checksum covering all data structure members including the data value elements and

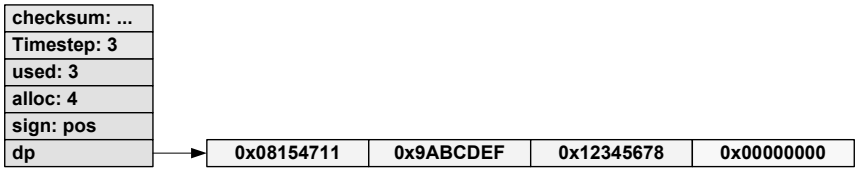


Fig. 11: Software-based implementation of SafeBigNums

- a Timestep structure member, which reflects the SafeBigNum's discrete Timestep.

The software libraries must be extended in a way, that they implement the following functionality:

- verification of the operands' checksums prior to processing or using the data contained,
- setting or updating the checksum during data processing,
- verification of the operands' Timesteps prior to processing or using the data contained and
- setting or updating the Timestep values during data processing.

References

[1] AEG Datenverarbeitung: TR 4 Bedienungshandbuch
[2] A. Bradbury, G. Ferris, R. Mullins: Tagged memory and minion cores in the lowRISC SoC; 2014; `http://www.lowrisc.org/docs/`
[3] D. T. Brown: Error Detecting and Correcting Binary Codes for Arithmetic Operations; IRE Transactions on Electronic Computers; Vol. EC-9, Issue 3; 1960
[4] U. Dhawan, N. Vasilakis, R. Rubin, S. Chiricescu, J. M. Smith, T. F. Knight, Jr., B. C. Pierce, A. DeHon: PUMP: A Programmable Unit for Metadata Processing; `http://www.crash-safe.org/docs/PUMP-HASP-2014.html`; 2014
[5] P. Forin: Vital Coded Microprocessor Principles and Application for Various Transit Systems; IFAC Control, Computers, Communications; S. 79–84; 1989
[6] I. Koren, C. Krishna: Fault-Tolerant Systems; 1. Auflage, 2007; Morgan Kaufmann Verlag; ISBN 978" 0" 12" 088525" 1

[7] A. Mayr: The Architecture of the Burroughs B5000 - 20 Years Later and Still Ahead of the Times?; 1982; http://www.smecc.org/The%20Architecture%20%20of%20the%20Burroughs%20B-5000.htm
[8] U. Schiffel: Hardware Error Detection Using AN-Codes; PhD thesis; Technische Universität Dresden; 2011
[9] T. St Denis, G. Rose: BigNum Math; 2006; SYNGRESS; ISBN 9-781597-491129
[10] ST: STM32F0 Series safety manual; UM1741 - Rev 7
[11] S. Widmann: Eine inhärent sichere Mikroprozessorarchitektur; 2015; VDI Verlag; ISBN 978″ 3″ 18″ 384310″ 7
[12] S. Widmann: Eine Datenspezifikationsarchitektur – Methoden zur Datenflussüberwachung in sicherheitsgerichteten Echtzeitsystemen; Dissertation; FernUniversität in Hagen; 2017
[13] Multiple authors: List of arbitrary-precision arithmetic software; https://en.wikipedia.org/wiki/List_of_arbitrary-precision_arithmetic_software

Evaluation of Hybrid Measurement Structures

Gerhard Sartorius[1] and Steven R. Talbot[2]

[1]Chair of Computer Engineering, FernUniversität in Hagen, Germany

[2]Institute for Laboratory Animal Science, Hannover Medical School, Hannover, Germany

Abstract: In hybrid-measurement structures, e. g. classification, processes must often be carried out thousands of times per second to see if a match for an object is detected. Therefore, it is necessary to combine fast measuring methods with slower but more accurate ones to balance effectivity and precision in form of a staggered measuring for the realization of practical solutions. For a safe and effective operation of such hybrid structures we propose a command line interpreter to create the necessary commands for controlling the overall procedures.

1 Control of Complex Measuring Tasks

Command-line interpreter instructions which are to be managed and applied in the form of a list for specific measurement objects or for measuring tasks are helpful for generating evaluation procedures in complex measurement systems (hybrid-measurement structures). This can be done without changing the code of the overall measurement system. A prerequisite for doing this is the organization of the measurement system into suitable functional sets (modules). They must be independent of each other, individually creatable, testable and usable.

2 Realization Concept

A computer unit (e. g. a tablet) for displaying results, controlling measuring procedures and entering command-line lists (batch-files) which will function in combination with a hybrid-measurement structure, consisting of measurement channels and modules, is envisaged. This hybrid measuring structure (realized on a PLC or FPGA basis) performs the measuring and processing (MCM). In figure 1 a corresponding arrangement for this is shown. The computer unit (OMDU) interacts with the coordination unit, in which the command-line interpreter is integrated. An example with interpreter commands is given in tables 5

and 4. The command-line interpreter for converting the command lines into directives for controlling the measuring process is integrated in the coordination unit of the target system. Information for the command-line interpreter is listed in table 4 and a example is given in section 4.2. Because of the channel-orientated and modular design, modules can be selected and linked with interpreter commands. The results can be displayed in different ways. For a given application, the list of command instructions can be varied in the setup phase (SP) until the optimal processing scenario is found. With the last final settings of command-line list, the measuring system is adjusted and prepared for the measurement application in the work phase (WP). All this can be achieved without changing the FPGA programming.

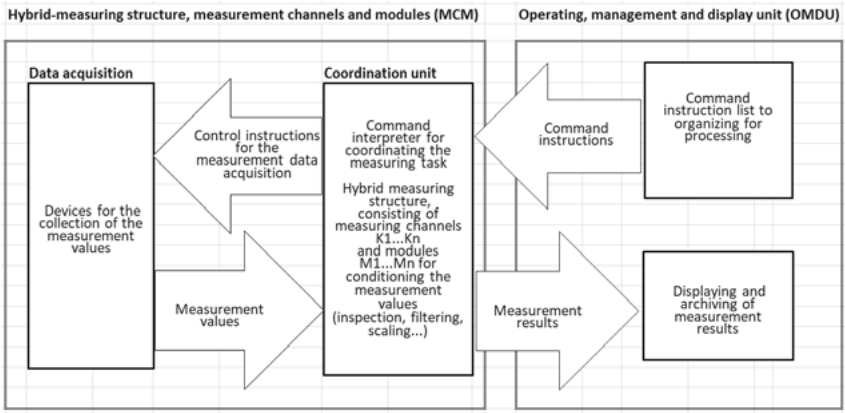


Fig. 1: Command interpreter to organize the measured value processing

3 Hybrid Measuring Structures

With complex measurement tasks, thousands of measurements per second must be carried out to determine if measurement values for a specified characteristic or significant measuring values in a defined area for an object to be analyzed (e. g. inspecting the surface of an object) are present. Such measurements may be used, for example, in quality-assurance inspections. For these measurements it is important to identify relevant areas in a very short time. An example for a hybrid measurement-structure is shown in figure 1. The modules for data acquisition, processing, scaling and evaluation can be realized, for example, with an

FPGA device. The **data acquistion** and the **coordination unit** together form the device **measurement channels and modules** (MCM) in figure 1. The **coordination unit** communicates with the **operating, management and display** device (OMDU). For this it is advantageous to combine rapid measuring methods (designed for maximum speed) with slower methods for higher accuracy (designed for high precision) to balance effectiveness and precision in line with good practice and in the form of a staggered measurement procedure [1].

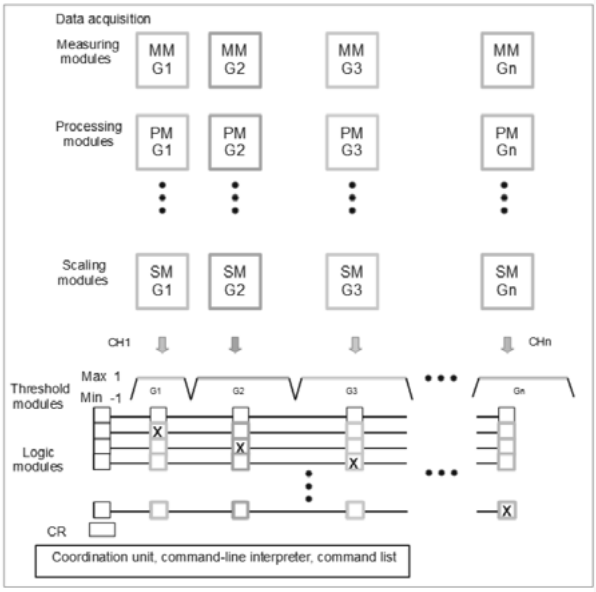


Fig. 2: Data acquistion, Coordination unit and specific modules for measuring, processing, scaling and thresholding, controlled by the command-line interpreter

For Calculations of lower complexity, the control of the measurement process, the selection and coordination of the individual modules and channels and the communication with the superior unit OMDU (figure 1) and the functionality of the command-line interpreter can be carried out by a CPU, which is implemented in the FPGA-device [4]. Figure 2 shows a possible structure for the measurement results of the measuring channels (CH1, ..., CHn) and measuring modules (MM1, ..., MMn) for the acquisition of the measurement data, modules for controlling the measurement process (PM1, ..., PMn) in combination

with scanners, spectrometers..., further process modules (PM, ...) for processing, transformation (e. g. into a representation with coefficients), filtering..., scaling of the measurement results, generation of Boolean expressions with the **combination result** (CR), e. g. whether the measurement values for a measured object are in the tolerance band or not. The **coordination unit** controls each individual module via the implemented command-line interpreter with the list of interpreter commands transferred and stored in the **coordination unit**. A staggered measurement includes (for example) a rapid measuring method in combination with a slower method for higher precision for evaluating the areas of interest, which have been identified by the rapid measuring method. Within the rapid measuring method, the coordinates of the areas of interest will be recorded. Within these areas, a method of high precision can be used to confirm the correct results and to reject the inaccurate ones. In this way shorter overall measurement times will be achieved. This permits a more efficient examination of larger surfaces by using an appropriate scanning device [3]. Furthermore, for a given measured object, different measuring methods can be used, and the best method can be selected for the respective measuring task. This and other combinations can be created in a well structured and efficient manner with arrangements as shown in figure 2, in combination with interpreter commands as introduced in chapter 2. With these commands a command sequence (command program) for controlling the modules of the measurement system can be created. The **hybrid measuring structure** allows relationships between measuring channels and modules. This helps to balance measuring speed and measuring precision in a practical manner.

4 Command-line Interpreter and Interpreter Commands

In this section different interpreter commands and associated information for setting up the **coordination unit** to adjust the measurement system will be shown. The measurement results and the settings can be displayed in an evaluation window and registered in a data-file. Threshold results of different channels can be combined to a **combined result** (CR). It is to be noted that the results of different measuring channels must be true, e. g. CR = xV.CH1 AND xV.CH2 AND xN.CH3 in order to be certain, especially when the impacts of the behavior of measuring devices and/or disturbances are considered. The description of the syntax is listed in tables 1 and 3. In general, the instruction set is intentionally restricted to a minimum to make the measurement system an efficient and easily understandable tool for the user. As an example of this principle, operands and operators were combined into a single expression for the measuring

system in figure 1. The command-line interpreter must be integrated in the target system (MCM). The resulting command program in the **coordination unit** consists of a list with instructions which is transferred from the OMDU-device to the MCM-device in the **coordination unit**. For a specific measurement system, a catalog lists all instruction set shows all instructions to the user. In the simplest case, the instructions are separated by a space character in the command line of an appropriate terminal or PC-program. Pushing the **enter button** causes the transfer of the characters to the target system. The processing of the transferred instructions is executed by the command-line interpreter[1]. In a more complex case, the instruction list is stored in an instruction-file (similar to a batch-file). The file can be transferred from the command line into the target system. They can be activated directly or by an another instruction-file. The instructions were processed by the command-line interpreter. To follow the instructions for performing an evaluation scenario, the instruction-file is loaded into the target system as a txt-file (WRITE<*file*>), and after this, the processing starts with the start-command (START<*file*>). This design permits rapid changes for requirements during the evaluation or test of the system. When the instructions for the evaluation are optimized, the instruction set is used for a special measuring task and can be stored and managed in the *Operating, management and display unit* (OMDU). The unit **management channel and modules** (MCM) runs with the last instruction set which has been transferred into the **coordination unit**. With the address numbers of the operands (index) for adjusting the module parameters in the measurement system (table 2), the data direction and a predefined functionality is automatically set. The same applies to operands for the reading of process values in table 1. Declared markers, which can be used in the instruction list (command program), are listed in table 1. If additional functions are needed, the list of operands in the target system can be expanded by an appropriate command. Each SDO or PDO command is sent to the target system in the form of an 8-byte string and can be handled in an instruction list, to be called up in the target system. Examples for this are listed below at appropriate places. The coding of instructions and operation corresponds to the numbers in column 2 in table 2 and table 3.

[1]The way commands are processed is closely related to the CANopen standard.
CANopen is mainly used as an embedded network for machine control and for linking control units in the automobile industry and is used for general industrial control systems for short distance-based devices in accordance with ISO 11898-1 (CAN-protocol) and ISO 11898-2 (fast physical transfer layer). The CANopen application layer is normed by CENELEC as EN 50325-4 [2].

4.1 Instruction Types and Module Parameters for Setting up a Functionality

Concept main-sub instruction: with the main instructions for specifying the module and the subinstructions for the fine adjustment of the modules behavior, there are, theoretically, 65536 different possible settings for each processing area (MM, PM, ..., SM, CR).
Operand: implied type declaration for process values, parameters and constants. Declaration of markers (placeholders for received values and other values), for sending parameters and for reading results and messages.
Definitions: setup phase (SP), work phase (WP) in table 2 and 3.
Main instruction SP, WP: in table 2 and 3, coding numbers (Ix) in columns 4,6,8,10 (D1, D3, D5, D7).
Sub instruction SP, WP: in table 2 and 3, coding numbers (Ix) in columns 5, 7, 9, 11 (D2, D4, D6, D8).
Enter button: submits a command line, represented by <CRLF>.
Data types: binary is represented by BOOL, an whole number by INTEGER, a floating-point number by REAL[2].
xy: placeholder, stands for a setting which is based on the main function, whose functionality is stored in the target system.
Index: With the index, the respective instruction is defined or referenced. The index can be interpreted by the target system, and, furthermore, the direction of the data flow is defined by it. These parameters are predefined and stored on a list in the target system. They can be changed via instructions or via command line. The operand with channel 0 resets the referenced block to the preset condition of the stored list.
Example for a reset transmission: ulP.5000 0 0 0 0 0 0 0 0 <CRLF>
for resetting the modules in the work phase to default settings.

4.2 Operands

Receive: To be able to receive different formats, the last four bytes of the PDO telegram in table 3 are reserved during read-out operation for selection purposes. With the index (identifier), where and in which direction the data is to be sent is defined (producer-, consumer principle) [2]. For the xV-format 8 bits are needed, and for the ulV-Format 32 bits required. The 4 bytes correspond to the Big Endian order HH, HL, LH, LL and are stored in the string. For the xV-format 8 bits are needed, and for the ulV-format 32 bits required. The 4 bytes correspond to the Big Endian order and are stored in the string. After reading

[2]REAL is a 32 bit floating point as defined by IEEE 754.

Table 1: Declaration of the variables for a command program: after the point, follows the index (described in tables 2 and 3) for SDO and PDO operands, for marker-operands, an identifier must be specified, Index (Ix), Channel (CH), degree of membership (DM), variable (V), function (FN)

No.	Representation	Data type	Application	Syntax
1	xV.CH1	BOOL	Process value	
2	iV.FNCH1,	INT (32 bit)	Process value	
3	ulV.3001	QINT (64 bit)	Receive process value	ulV.3001 <CRLF>
4	rV.DMCH1	REAL (32 bit)	Process value	
5	xP.12	BOOL	Parameter	
6	iP.3	DINT (32 bit)	Parameter	
7	ulP.4001	QINT (64 bit)	Module parameter, transmit	ulP.4001 4 8 10 1 16 1 30 1 <CRLF>
8	ulP.2001	QINT (64 bit)	Module parameter, receive	ulP.2001 4 8 10 1 16 1 30 1 <CRLF>
9	rP.254	REAL (32 bit)	Parameter	
10	TRUE	BOOL	Constant	
11	-5439	INT (32 bit)	Constant	
12	-89,76	REAL (32 bit)	Constant	

the telegram, the command-line interpreter transfers the data automatically to the appropriate marker, which is specified in the operand.

Examples for Receiving Process and Parameter Values

Process value work phase: ulV.3001 <CRLF>. Receive measurement results from channel 1. The data is cyclically transferred to the target and will be stored in the associated marker xV.CH1. With this marker, the data can be further processed. Parameter values set up phase: ulP.2001 4 8 10 1 16 1 30 1 <CRLF>. Receive the module settings from channel 1.

Table 2: Operands and module parameters, service-data objects (SDO), process-data objects (PDO) for setup- and workphase, dataword D1 (D1) ... dataword D8 (D8), main-instruction data processing (DP), sub-instruction data processing (dp), main-instruction X space (X.S), sub instruction x (x.s) ... up to sub instruction z (z.s); in column 3 ... 11: capital letters for main-instructions (.S), lower-case letters for sub-instructions (.s), index (Ix), channel (CH), degree of membership (DM), variable (V)

(SDO) SP transmit	Ix		D1	D2	D3	D4	D5	D6	D7	D8
Operand	Ix	CH	DP	dp	X.M	x.s	Y.M	y.s	Z.M	z.s
uIP.4000	4000...	0	0	0	0	0	0	0	0	0
uIP.400n	400n	n	4	1	11	1	18	1	30	1
(SDO) WP transmit	Ix		D1	D2	D3	D4	D5	D6	D7	D8
Operand	Ix	CH	DP	dp	X	x	Y	y	Z	z
uIP.5000	5000...	0	0	0	0	0	0	0	0	0
uIP.500n	500n	n	4	1	8	1	18	1	29	1

Table 3: Operands for reading of process values from the measurement system, capital letters for main-instructions (.S), lower-case letters for sub-instructions (.s); index (Ix), channel (CH), degree of membership (DM), variable (V)

(SDO) SP receive			D1	D2	D3	D4	D5	D6	D7	D8
Operand	Ix	CH	DP	dp	X	x	Y	y	Z	z
ulV.2001	2001 ...	1	x	x	x	x	x	x	x	x
ulV.200n	200n	4	x	x	x	x	x	x	x	x
(SDO) WP transmit			D1	D2	D3	D4	D5	D6	D7	D8
Operand	Ix	Ch	DP	dp	X	x	Y	y	Z	CHi.LL
ulV.3001	3001...	1	x	x	x	x	x	x	x	CH1.LL
ulV.300n	300n	4	x	x	x	x	x	x	x	CHn.LL
Operand	Ix	CH	DP	dp	X	x	FNCHi.HH	...	...	FNCHi.LL
ulV.3101	3101...	1	x	x	x	x	FNCH1.HH	...	...	FNCH1.LL
ulV.310n	310n	4	x	x	x	x	FNCHn.HH	...	...	FNCHn.LL
ulV.3201	3201...	1	x	x	x	x	DMCH1.HH	...	...	DMCH1.LL
ulV.320n	320n	4	x	x	x	x	DMCHn.HH	...	...	DMCHn.LL

Operators and Instructions for Command Programs and Subprograms

Interpreter instructions for the writing of programs, which will be transferred as a textfile (.txt-file) and are to be read out and evaluated by the command-line interpreter, are listed table 4. For more information refer to the CANopen standard ISO 11898-1 (CAN-protocol) and ISO 11898-2 and in [2].

Example for an Instruction List of a Batch File (.txt-file):
%Name; ReadMeasuring.txt
%Function: read result of CH1 ...CHn, create **combination result** CR
ulV.3001 <CRLF> % start cyclic measuring channel 1, register in xV.CH1 ...
ulV.300n <CRLF> % start cyclic measuring channel n, register in xV.CHn
CR := xV.CH1 AND xV.CH2 AND xV.CH3 AND ...xV.CHn
% CH1 ...CHn = TRUE => match

Table 4: Interpreter instructions

Syntax	Function
%	Comment marker: after % the command line will not be interpreted any further. Space for comments.
A:=	Assignment to variables: variable A:= value, variable A:= CH1 (result of channel 1)
#mark:	Mark: is a placeholder for the text of a specified destination, #=start textmark, := end textmark. Example: jump with instruction goto *End* to #End:
repeat	Begin of a repeat loop,
...	place for instructions,
until <condition>	Jump to the previous repeat mark, if <condition> = 0
...	place for instructions,
goto <mark>	Continues the program execution at the location #mark:
gosub <mark>	Continues the program execution at the location #mark: (start of the subprogram)
return	Must be the last instruction in the subprogram, continues the program execution after the last executed gosub-instruction
print(<position>, <Operand>)	Prints text and values in the message window or into a file. After printing, a line feed will be carried out. position: line number in the message field
print(<position>, <operand>;)	The semicolon suppresses the line feed
WRITE<*file*>	Textfile xy.txt is written into the target system
READ<*file*>	Textfile xy.txt will be read out of the target system
START<*file*>	Textfile xy.txt will be started in the target system
STOPP<*file*>	Textfile xy.txt will be stopped in the target system

Table 5: List of operators

No.	Operator	Function	BOOL	INT	REAL
1	OR	Logical OR	x		
2	AND	Logical AND	x		
3	XOR	Logical exclusive OR	x		
4	NOT	Logical negation	x	x	x
5	$+$	Addition		x	x
6	$-$	Subtraction		x	x
7	$\cdot$	Multiplication		x	x
8	$/$	Division		x	x
9	$<$	Less than		x	x
10	$<=$	Less than or equal to		x	x
11	$>=$	Greater than or equal to		x	x
12	$>$	Greater than		x	x
13	$=$	Equal	x	x	x
14	$<>$	Unequal	x	x	x

References

[1] Ahlers, Horst; Waldmann, Jürgen: *Mikroelektronische Sensoren* Berlin, Heidelberg Springer-Verlag (1997)

[2] Koppe, Uwe: *AN1201 Einführung in CANopen* Internet (2018) (`http://www.microcontrol.net/download/appnotes/canopen_intro.pdf`)

[3] Mandal, Oliver: *Erstellung und Validierung einer kombinierten NIR/Raman-Bibliothek zur Identitätskontrolle organischer und anorganischer Substanzen* Internet (2009) (`https://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-5143/mandal_gesamt.pdf`)

[4] Wakerly, John F.: *Digital Design Principles and Practices* Upper Saddl River, New Jersey Prentice Hall (2001)

Image Processing

A Comparative Study of Visual and Visual-inertial SLAM

Xin Xu[1,2], Chunrong Yuan[1] and Zhong Li[2]

[1]Autonomous Systems Lab
Faculty for Information, Media and Electrical Engineering
TH Köln – University of Applied Sciences, Cologne, Germany

[2]Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

A major problem in the field of mobile autonomous systems is the determination of the positions of a moving robot in an unknown environment. With its root in probabilistic state estimation, the SLAM (simultaneous localization and mapping) approach combines the interconnected processes of environmental sensing, robot localization and map building altogether. Capable of incorporating the uncertainties inherent in the robot movement and sensor measurement process, a robot facilitated with proper sensors can explore an unknown place, reconstruct its whole movement trajectories and at the same time build a map with the locations of 3D points that belong to objects in the environment.

For outdoor environments, one can use GPS for position measurement. In indoor environments, visual information can be extracted for distance estimation. Typical visual sensors are single or stereo cameras, depth cameras, LiDAR (Light Detection and Ranging) etc. Combined with additional sensors such as IMU (Inertial Measurement Unit) or wheel encoders, it is theoretically possible to achieve more accurate localization and mapping performances.

The goal of this ongoing work is to develop novel computer vision approaches, which can be used on different robot platforms for the purpose of map building and robot navigation. Particularly, we are interested in using monocular and stereo cameras, as one can use the captured images for the recognition of individual objects so that the final map contains not only position parameters but also semantic information about the explored environments. By using statistical and especially deep-learning techniques, semantic maps can be built and used for intelligent vision-based navigation of robots.

As first step, different visual SLAM approaches so far available have been studied. Our survey points to two state-of-the-art algorithms: ORB-SLAM2 [1] and VINS-SLAM [2]. Both can be used with either monocular or stereo cameras.

The later can also be fused with IMU data [3]. Following the literature review, we have carried out performance evaluations of the two SLAM approaches under different sensor combinations. For benchmark purpose, the experiments have been carried out using camera and IMU sensor data recorded in the EuRoC dataset, which is publicly available at `https://projects.asl.ethz.ch/datasets`. Using different algorithms under different sensor configurations, the 3D positions and movement trajectories of the robot have been calculated. The results achieved in this comparative study are shown in Fig 1.

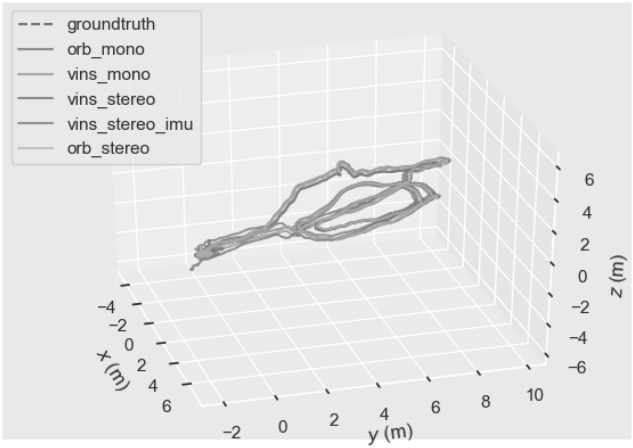


Fig. 1: Robot positions calculated with different algorithms and sensor configurations.

References

[1] Mur-Artal, R., Tardós, J. D.: ORB-SLAM2: An open-source SLAM system for monocular, stereo and RGB-D cameras, *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017

[2] Qin, T., Li, P., Yang, Z., Shen, S.: VINS-Mono: A robust and versatile monocular visual-inertial state estimator, *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018

[3] Qin, T., Cao, S., Pan, J., Shen, S.: A general optimization-based framework for global pose estimation with multiple sensors, *arXiv: 1901.03642*, 2019

Design of a Small-scale Ground Vehicle for Autonomous Navigation

Maximilian A. De Muirier[1], Chunrong Yuan[1] and Zhong Li[2]

[1]Autonomous Systems Lab
Faculty for Information, Media and Electrical Engineering
TH Köln – University of Applied Sciences, Cologne, Germany

[2]Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

The fields of mobile robotics, especially that of autonomous vehicles, has been experiencing recently a rapidly growing interest. Due to its broader availability and decreasing acquisition costs, robotic technology is becoming more and more affordable for both academic and industrial use. This work deals with the complete process of the design and development of a novel autonomous ground vehicle (AGV). The AGV is designed as a small-scale robotic platform for autonomous navigation, where three Raspberry Pi 3 Model B serve as computing units for sensing, communication and course control during navigation. While the main focus is set on a simplistic and holistic approach towards robotic design and implementation, special attentions have been paid to various factors including robustness, precision, energy efficiency, behaviour predictability, system extensibility as well as reproducibility. The platform has a relatively small footprint and yet sufficient place for holding several kinds of sensors including 2D and 3D cameras, IMU (Inertial Measurement Unit), ultrasonic and LiDAR (Light Detection And Ranging), with which autonomous exploration of unknown environments can be achieved. As an elegant, sophisticated and versatile platform, the robot can be used for the purpose of both research and education.

A Comprehensive Analysis of Document-Image Distortions and their Respective Impact on Distorted Text/Character-Image Recognition Quality

Perumadura De Silva, Kolli Abhiram, Vahid Tavakkoli,
Kabeh Mohsenzadegan, Kyandoghere Kyamakya,
and Jean Chamberlain Chedjou

Institute of Smart Systems Technologies, Transportation Informatics Group
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

Ralf Pichler, Olaf Bouwmeester, and Robert Zupan

UNiQUARE Software Development GmbH, Krumpendorf, Austria

Abstract: This paper presents a brief comprehensive discussion of the impact of the presence of document-image distortions on the recognition performance of an Optical Character Recognizer (OCR). First, the paper introduces modern issues related to document text extraction, specifically in the presence of natural distortions such as noise, blur and contrast. Then, the discussion continues by introducing the related mathematical approximations of the distortions. Lastly, the performance of three simple classifier-based OCR system mini-prototypes is evaluated through a comprehensive sensitivity analysis, whereby the obtained sensitivity curves do show and represent the impact of respective distortion levels on the OCR accuracy. The mini OCR protypes designed and used in this paper are three neural network architectures: a simple two-layer Convolutional Neural Network (ConvNet), an extended more complex ConvNet-Plus, and an Echo-State Cellular Neural Network (ES-CellNet). Further, the dataset used for training and testing is MNIST. The training contains only distortion free images; but the testing dataset is a mixture of distortion fee and distorted images.

An image enhancement model based on ES-CellNet is designed to enable an observation of the improvement in OCR accuracy when the distorted images are pre-enhanced. The benchmarking results shows, overall, that the ES-CellNet performs better than the ConvNet and almost as good as ConvNet-Plus at the OCR (i. e. character recognition)

task. Both ES-CellNet and ConvNet-Plus show a significant character recognition improvement when involving an image-enhancement pre-processed input character image; this does justify the requirement of a preprocessor for improving the OCR accuracy.

1 Introduction

The digitalization of offices (with the positive consequence of "no more hard-copy documents" in office related processes) is a crucial component in the modern business world where it allows businesses to automatize both data extraction (from digitalized hard copy documents) and management processes [1, 2]. This leads to a better management of information and helps a business to minimize the utilization of manpower [3]. Additionally, as a result of office digitalization, the consumption of hardcopy-paper documents can be eliminated, which does also partly lead to a better space utilization in the offices [3]. A major issue, amongst others, in office digitalization is the data/text extraction from the document-images. This does generally involve an Optical Character Recognition (OCR) system [2, 4]. A document-image can generally be acquired (i. e. digitalization) through either scanners (traditionally) or (more recently becoming a more popular alternative) mobile phone cameras. The acquisition process of document-images generally introduces several distortions such as noise, blur, contrast, and illumination variations [5, 6]. These distortions do generally significantly degrade the quality of the text-data extraction through OCR [6], leading to a consumption of manpower in order to manually correct the false/imperfect OCR readings. Therefore, a system that acts as a preprocessor is required to remove the distortions and enhance the quality of the acquired images prior to the OCR involvement.

It is generally difficult to enhance a document-image w.r.t. the named distortions, especially regarding blur [6] and some especially non-uniform illumination differences. The document-image enhancer does at times introduce artefacts or deformations which, despite some apparent relative quality improvement, do still not enable a significant OCR related readability improvement [7]. The aftermath of the lastly mentioned character deformations and/or artefacts (possibly due a non-robust enhancement) is the false recognition of some characters or even complete words by the OCR system [7].

1.1 A Comprehensive Overview of the Research Questions Addressed

Besides providing a comprehensive review of the various observed distortions and their respective origins, this paper does also develop and validate a novel document-image enhancement concept.

Specifically, in this paper we do address the following research questions (RQs) and objectives:

- RQ-1: What are the naturally introduced distortions in a document-image and how do they affect the related OCR performance? Indeed, a document image can be contaminated by either one single or a mixture of various distortions such as noise, blur, geometric distortions, and paper-aging induced distortions, which do potentially degrade the OCR performance. These distortions can approximatively be modelled mathematically in order to partly mimic their natural behavior. Then, this last named mathematical approximation can be used to generate a novel dataset to be used for *appropriately training document-image enhancement* neural network models. More details on distortions modeling and the related OCR performance are presented and comprehensively discussed in Section 2.
- RQ-2: How different distortion levels do impact the OCR classification/ detection accuracy and what are the feasible ranges of distortions? The impact of the individual distortion levels on the performance of an OCR system is an important factor to observe. The current related literature does not provide yet a systematic study of the question. This shall give an insight to the acceptable level of a particular distortion type as well as which distortion type affects the OCR performance more. In order to evaluate the impact of the individual distortions, a related sensitivity curve can be produced from the classification results (of the OCR) against the respective distortion classes/levels. A comprehensive discussion of the conducted sensitivity analysis is given in the Section 3.

1.2 Brief Overview of the Paper

The rest of the paper is structured as follows. In Section 2 we address RQ-1 by discussing the document-image distortions, their mathematical modelling and the respective impact on OCR recognition performance. The RQ-2 is discussed in the Section 3 by presenting a sensitivity curves of the MNIST classifiers along with their benchmarking and the discussion of the results. Finally, concluding remarks are presented in Section 4.

2 Comprehensive Review and Analysis of Most Common Naturally Induced Distortions and How They Can Affect OCR Performance (RQ-1)

2.1 Global Categorization of the Distortions w.r.t. the Respective Origins

In order to enhance an acquired document-image for improving the quality of the OCR results, it is important to understand and organize the distortions induced through the image acquisition process. In the relevant literature, there have been several attempts to identify and categorize the individual distortions that could affect the OCR performance [8–10].

The distortions in a document-image can be organized according to its natural context of occurrence or origin. One type of occurrence of the distortions is due to the bad external environmental and lightning conditions. Examples of these distortions are spotlight, illumination variations, and shadow casted on a document-image [11]. Another main source that induces distortions such as noise [11] in an image are impurities induced by camera hardware. Then further distortions under the blur label are those such as focus, bilateral and motion blur [10] can be explained as the distortions due camera manipulation by a human or machine operator. When acquiring a document-image through mobile phone cameras, the orientation of the document introduces geometric distortions which could also possibly potentially lead to poor OCR performance, example due to changes in the orientation of certain characters [12]. Finally, in the case of historical documents, apart from the already mentioned most common distortions such as noise and blur there exists degradations due to aging of the hard physical paper of the document [13].

Whenever these above listed distortions are present (individually or in groups) within the characters of a document-image, the infected characters will appear and thus be represented differently to the OCR system. As a result, the OCR system will either perform a false detection of those deformed characters or there will be no detection at all.

2.2 An Introduction to the Internal Dynamics of an OCR

An optical character recognition system plays the main role in text data extraction from document-images. Similarly to many image processing applications, an OCR system consists of preprocessing, OCR and postprocessing modules [14]. Figure 1 presents a high-level overview of an OCR system.

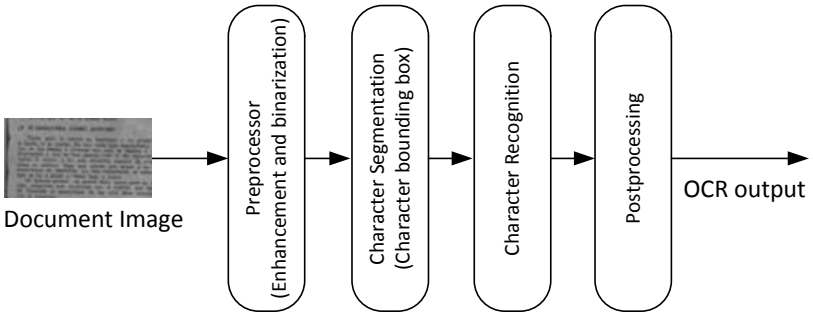


Fig. 1: A high-level overview of an OCR system

The preprocessing module performs a series of necessary low-level operations such as image denoising, binarization and line removal from a document-image [15]. These preprocessing steps are designed and implemented in order to minimize the stress on the text extraction process. The steps such as noise and line removal help the OCR to extract the text accurately by removing unnecessary artifacts such as noisy pixels and information such as tables. On the other hand, the binarization process of a document-image allows an OCR to differentiate between foreground and background properly, allowing the OCR system to process images faster.

Once the preprocessing is completed, the preprocessed document-image is given to the image processing module of the OCR. This module performs the tasks such as segmentation and text extraction [16]. The segmentation method constructs a bounding box around the detected characters and the text extractor recognizes the character within a given character segment.

Depending on the method used for text extraction, an OCR can be categorized into two categories.

Category 1: Matrix/Pattern matching: This method uses a set of known glyphs to recognize a character. The recognition is achieved by computing the pixel similarity between a character segment and glyphs [17].

Category 2: Feature extraction: The idea behind this method is to extract the features that construct a particular character. Once the features are properly extracted, a classifier can utilize these features to recognize the character within a character segment [17].

The detected characters from the OCR part of the system should be organized accordingly to make sense for a human-user. Therefore, the main task of the postprocessor is to construct the words from the detected characters. Additionally, some postprocessing modules can perform a lexicon restriction [18]. In the lexicon restriction the postprocessor search for the similarity between the detected word and the words in a dictionary. When the detected word is not exactly similar to the one found in the dictionary the detected word is replaced by the most similar word from the dictionary.

When processing a strongly distorted document-image, the off-the-shelf OCR systems (especially the free OCRs such as Tesseract [19]) usually fail. One of the main reasons for this failure is the lack of robust distortions-aware preprocessing of the document-images, since the current preprocessors of the (off-the-shelf) OCR systems are only tuned for a specific (probably rather very low) distortion level [11].

2.3 Mathematical Modelling of the Individual Distortions and Discussion of Realistic Values of the Respective Parameter Ranges

The impact of the natural distortions within the named various contextual origins can be observed better with an artificially generated dataset. Since the distortion levels are known/fixed, the generated distorted MNIST images can be categorized into 10 different classes for each distortion type. In order to generate these data, proper mathematical models that imitate the natural distortions are required.

In order to generate the distortions such as contrast, shadow, spotlight and brightness, which are categorized in group-1, following mathematical formula is used [20].

$$g(x) = \alpha f(x) + \beta \tag{1}$$

The parameter α is the gain coefficient which determines the contrast and the bias coefficient β expresses the brightness in a document-image. The output image with manipulated contrast is $g(x)$ and the input image is $f(x)$. These distortions are induced to a particular document-image by varying the α and β coefficients. Varying α while keeping β at 0 produces a contrast effect. For a shadow effect one can keep α at 0 and vary the β in a negative range. When the β range is positive while the α is 0, the brightness effect can be achieved. The contrast level applied to the MNIST dataset in this experiment is: $0 \leq \alpha \leq 0.8$ and $\beta = 0$.

On the other hand, the noise present in an image can be approximated as gaussian noise [21]. In order to induce gaussian noise the following formula can be used [22].

$$P(z) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-(z-\mu)^2}{2\sigma^2}} \tag{2}$$

This formula modifies the value of a pixel z according to a given gaussian distribution. The distribution is controlled by means of the parameters σ (the standard deviation) and μ, the mean.

The mean and standard deviation used in the MNIST data for this experiment the α lies between 0 and 0.1 and $\mu = 0$.

The blur that is induced due to the out of focus can be mathematically approximated as a gaussian blur [23] which is presented by the following formula,

$$D(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-z^2}{2\sigma^2}} \tag{3}$$

where, σ is the standard deviation and $D(x)$ is the generated gaussian distribution. The gaussian distribution generated by (3) can be used to construct a convolutional kernel that would induce out-of-focus blur to a particular image. This kernel has the highest weight at its center and the intensity of the weights gradually decreases when moving away from the center. Along with the constructed gaussian blur kernel, a blur image can be produced using the following formula [24].

$$g(i,j) = d(i,j) \otimes f(i,j) + n(i,j) \tag{4}$$

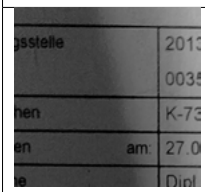
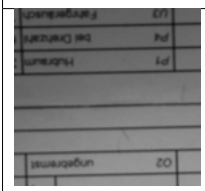
The i, j in the equation is the current center location in the image (row and column index), $g(i,j)$ is the produced blur image from the original image $f(i,j)$, $n(i,j)$ is the additive noise and $d(i,j)$ is the constructed blur kernel or the point spread function (PSF). The symbol $\otimes$ represents the convolution operation. For the Gaussian blur dataset, a standard deviation that varies from sigma $= 0$ to sigma $= 1.8$ is used.

2.4 Simultaneous Presence of Non-uniform Multiple Distortions on a Document Image and their Effect on the Character Recognition Performance

In a real-life scenario, a document-image can be contaminated with multiple and non-uniform distortions simultaneously. This situation occurs especially when a mobile phone camera is used for the acquisition of the document-image.

These non-uniform multiple distortions can be induced to a document-image by means of simultaneous individual distortions with varying parameter ranges. As an example, when generating an artificial document with non-uniform noise and shadow distortions, a randomly drawn σ for (2) and α for (1) is used along with the corresponding equations.

Table 1: Document-image examples (for illustration) containing non-uniform multiple distortions

Image examples	Contained distortions	Comments (e. g. explain briefly how the listed distortions may happen in reality)
	Shadow, noise, some low contrast	Eventually, non-uniform illumination of the document, etc.
	Spotlight, shadow, geometry distortion	Non-uniform illumination of the document and unreadable text because of the spotlight
	Out-of-focus blur, shadow	The presence of out-of-focus blur and slight shadow (bottom-right portion of the image) degrades the OCR performance

3 Neural Network Architectures, Performance Results and Discussion

3.1 Convolutional Neural Network Architectures

In order to evaluate the OCR performance of the distorted MNIST images, two ConvNet Classifier models as well as an ES-CellNet classifier are used. Each classifier has 10 outputs that corresponds to the 10 handwritten digits of the

MNIST dataset. Furthermore, an ES-CellNet image enhancer is designed to observe the classification (character recognition) performance once the distorted images are pre-enhanced.

First, a simple ConvNet consisting of 2 convolutional layers, 2 pooling layers and one classification layer is developed. The second ConvNet (ConvNet-Plus) is a more sophisticated model. It consists of two parallel convolutional modules that produce a concatenated feature set and two modules containing "layer normalization" and max-pooling. The ConvNet is trained with the original MNIST dataset and the ConvNet Plus is trained with the distorted MNIST dataset. The ConvNet Plus is trained on distorted MNIST data in order to observe the behavior of the classification accuracy when a ConvNet model is trained to classify MNIST digits in the presence of distortions. These two ConvNet architectures are shown in Figures 2 and 3 respectively.

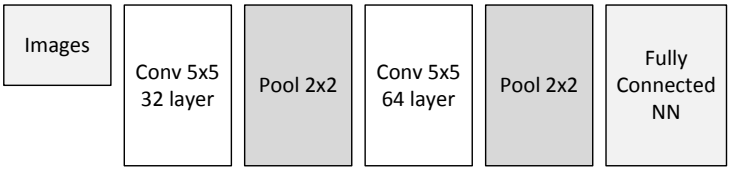


Fig. 2: A Two-layer ConvNet classifier designed to classify the distorted MNIST digits

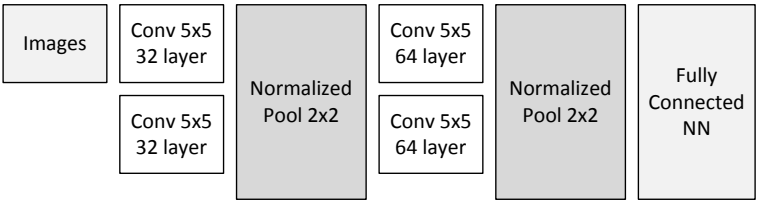


Fig. 3: The ConvNet-Plus classifier designed to classify the distorted MNIST digit classification

3.2 Echo-State CellNet Architectures

An alternative to the deep learning classifiers such as ConvNet and ConvNet-Plus is the Echo-State Neural Networks (ESNN) [25]. The ESNN comes under the so-called reservoir computing and it provides a very simple but strong neurocomputing methodology. An ESNN consists of three layers. They are the

input layer, reservoir and the output layer. The processing elements in the reservoir take the form of an RNN (recurrent neural) with leaky integration [26]. The mathematical model of the ESNN [26] is given below,

$$\tilde{x}(n) = \tan h(W^{in}[1 : u(n)] + W\, x(n-1)) \tag{5}$$

$$x(n) = (1-\alpha)\, x(n-1) + \alpha\, \tilde{x}(n) \tag{6}$$

where: x is the state, n is the time step, u is the input, W is the reservoir weights, W^{in} is the input weights and α is the leakage. The activation function used is $\tan h$. The state, input weights and reservoir weights are matrices and α is a scalar. The input is a matrix in most cases, but it can be also a scalar. In order to train the output layer of the ESNN, following weight update equation is used.

$$W^{out} = Y^{target} X^T (XX^T + \beta I)^{-1} \tag{7}$$

where: W^{out} is the output weights, Y^{target} is the target output, X is the design matrix and β is the regularization parameter. The regularization parameter is used to prevent any singularities in the matrix inversion. Every variable in this equation is a matrix except for the regularization parameter. The design matrix X is constructed by collecting the computed final states for each training input. This training method is called the Ridge Regression [26]. The following figure, i. e. Figure 4, presents the ES-NN model.

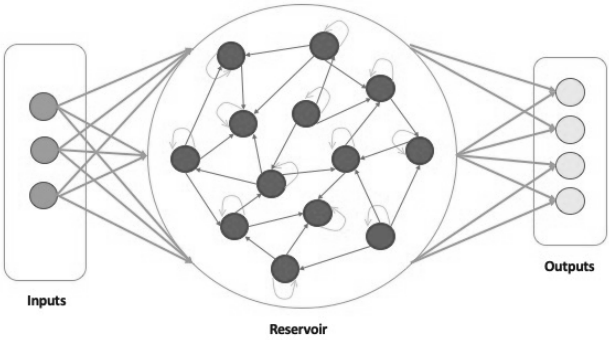


Fig. 4: The echo-state reservoir computing model

A novel substitute to the vanilla reservoir model represented in Eq. (5) and Eq. (6) is a Cellular Neural Networks [27]. Originally the CellNets are designed to function in the continuous time domain (CT-CellNet). But using existing

discretization methods, a CT (continuous-time) CellNet can be transformed into discrete-time CellNet (DT-CellNet). The equation of the DT-CellNet in general form is as follows, see Eq. (8) and Eq. (9) [28].

$$x(n) = A\,y(n-1) + B\,u(n) + I \tag{8}$$

$$y(n) = \tan h(x(n)) \tag{9}$$

where: x is the state, n is the time step, y is the cell output, I is the bias, A is the feedback template, B is the control template and $\tan h$ is the activation function. The final output of the model is taken when the cell reaches an empirically defined number of oscillations. The model diagram of the DT-CellNet is as follows (see Figure 5).

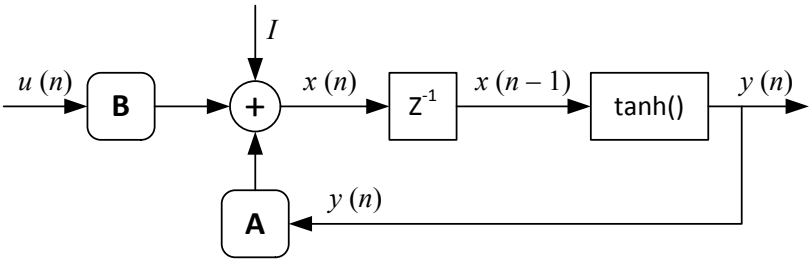


Fig. 5: Discrete time cellular neural network model

After replacing the processing elements of the reservoir with DT-CellNet, a two-layer ES-CellNet called ES-CellNet-1 was designed to perform the classification of the distorted images. The training process of this model takes following 6 steps,

Step-1: The training dataset that contains 700 original MNIST images is split into two where dataset-1 contains 300 images and dataset-2 contains all the 700 training images.

Step-2: A 28x28 MNIST training image is blocked into 16 blocks before giving as an input to the ES-CellNet-1 where each block has a size of 7x7. This blocking is done for all the training and testing images.

Step-3: The produced blocks are sequentially fed into reservoir 1 (Figure-6) and compute an output of size 160. This output layer represents the tanh activation of a particular class for a given 7x7 image block and since there are 16 image blocks and 10 MNIST classes, 160 output size is selected.

Step-4: The output layer 1 is trained with the training dataset-1, and once the training is done reservoir 1 and output layer 1 are set to inference mode.

Step-5: Then the output layer 2 (Figure 6) is trained with the dataset-2 using the first classification output (from output layer 1) as the input. The output layer 2 produces the final MNIST classification results.

Step-6: Finally, the rest of the distorted MNIST images (300 images) are used to test the ES-CellNet classifier.

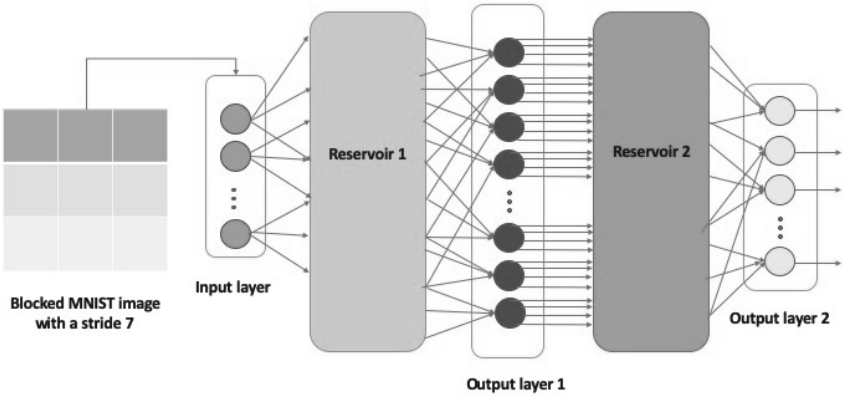


Fig. 6: ES-CellNet-1 classifier model. The output layer 1 produces the necessary classification features, where the output layer 2 produces the MNIST class predictions using the output layer 1 as the inputs.

The philosophy behind this two-step training process (see presence of and information gathered in output layer 1) is to capture the features missed by the reservoir 1 using the reservoir 2.

The functionality of the designed ConvNets and ES-CellNet is to realise a "Classifier" functionality which is similar to the text extraction step of an OCR which is performed after the character segmentation (as indicated in Figure 1).

Apart from the classification, another ES-CellNet called ES-CellNet-2 model is designed to perform the image enhancement in the presence of individual distortions. The training process of the ES-CellNet-2 enhancer model is as follows,

1. A set of 5x5 image blocks are extracted using a sliding window over a padded MNIST image with a stride of 1. This process is done to all the 400

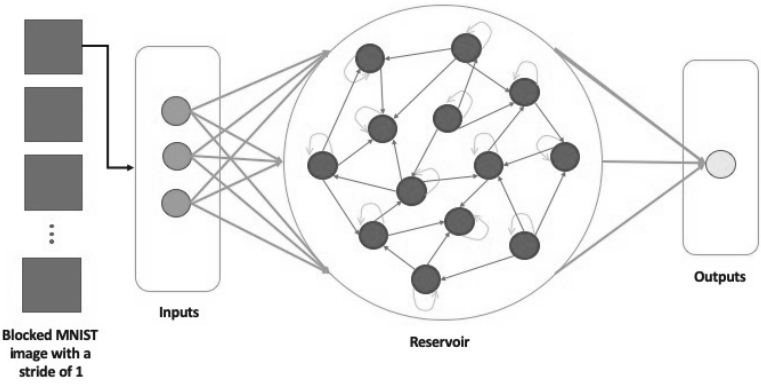


Fig. 7: ES-CellNet-2 based image enhancement model

training MNIST images and a new set of input images is created. The targets set is created such a way that for a given distorted input block, the target is the clean center pixel.

2. Using this new extended distorted MNIST dataset the ES-CellNet-2 shown in Figure 7 is trained to enhance the distorted MNIST images.

3. Finally, the rest of 600 MNIST images are used to test the model.

   - When enhancing one distorted MNIST image, the subjected image is blocked using a sliding window with stride of 1.
   - These blocks are processed using the ES-CellNet-2 and the corresponding enhanced center pixels are produced.
   - The enhanced center pixels are finally arranged to produce the final enhanced MNIST image.

3.3 Results: MNIST Classification in the Presence of Individual Distortions

In order to evaluate the sensitivity of these classifiers which are trained on original MNIST data, the distorted MNIST dataset is used. The sensitivity curve presents the accuracy of the classification w.r.t. the distortion classes. With the sensitivity curve, the impact of different distortion levels on the accuracy can be examined. The sensitivity curve of the ConvNet and the ES-CellNet classifiers in the presence of Gaussian Blur, Gaussian Noise and contrast are shown in Figure 8, Figure 9, and Figure 10 respectively.

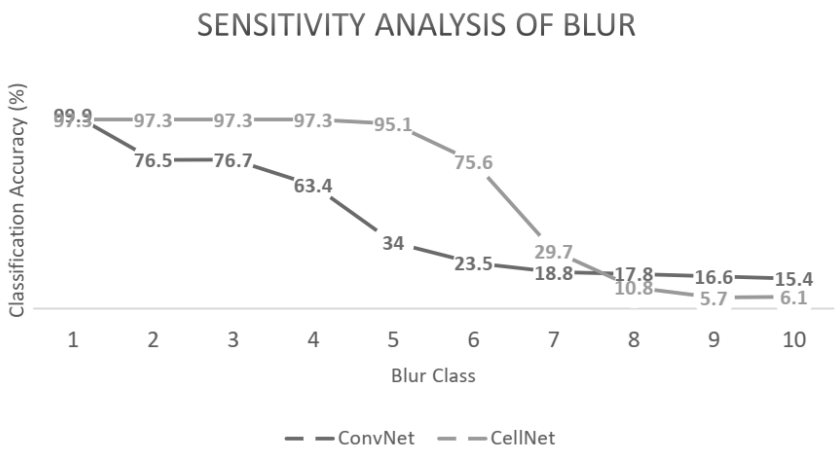


Fig. 8: Sensitivity analysis curve for the blur distortion. The x-axis represents the blur class of the image given to the MNIST digit classifier. Y-axis represent the accuracies corresponding to the blur classes.

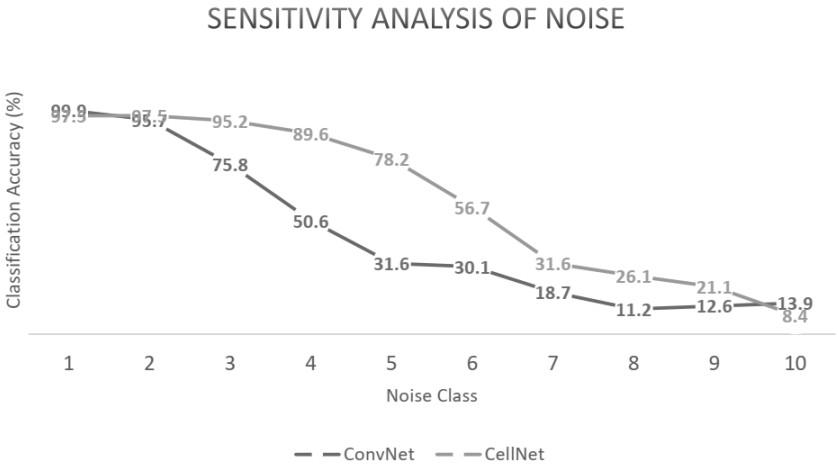


Fig. 9: Sensitivity analysis curve for the noise distortion. The x-axis represents the noise class of the image given to the MNIST digit classifier. Y-axis represent the accuracies corresponding to the noise classes.

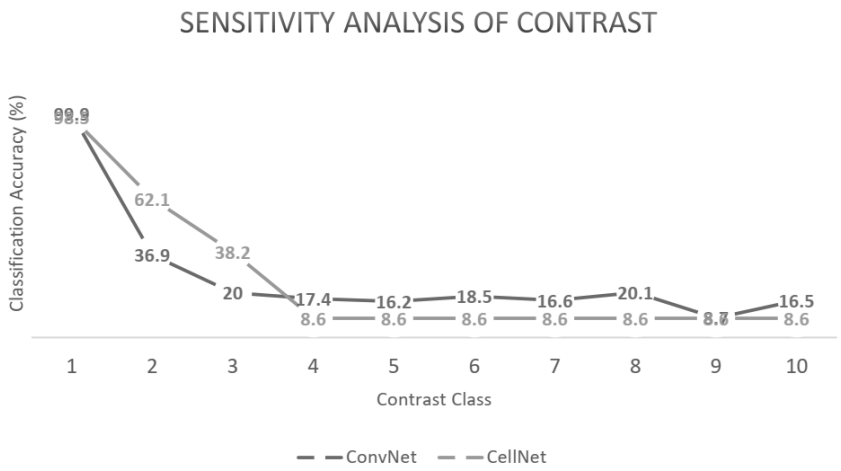


Fig. 10: Sensitivity analysis curve for the contrast distortion. The x-axis represents the contrast class of the image given to the MNIST digit classifier. Y-axis represent the accuracies corresponding to the contrast classes.

The average accuracies of the classifier models for all the three distortions are given in the following table (see Table 2).

Table 2: Average classification accuracies in the presence of different individual image distortions (input images are not enhanced)

	Average Accuracy values obtained		
Neural network Model	In case of Gaussian Blur	In case of Gaussian Noise	In case of Contrast
ConvNet	79.12 %	80.50 %	43.55 %
ConvNet Plus	95.50 %	94.16 %	44.16 %
ES-CellNet-1	97.30 %	94.97 %	51.80 %

A careful examination of the benchmarking presented in the Figure 8, Figure 9 and Figure 10 shows that the ES-CellNet-1 performs significantly better than the ConvNet for all the three distortions. For evaluation, a threshold of 70 % is defined as an acceptable level. In the case of Gaussian blur, the acceptable level of blur for ConvNet is class 3 where the accuracy is 76.7 %. But in the case of

ES-CellNet-1, the model maintains the robustness towards Gaussian blur until class 6 where the classification accuracy is 75.6 %.

When recognizing the MNIST digits in the presence of noise, similarly to the blur case, the ConvNet classifier can handle the noise up to class 3 with an accuracy of 75.8 %. On the other hand, unlike in the blur case the ES-CellNet can maintain an acceptable enhancement until class 5 with an accuracy of 78.2 %.

The Figure 10 shows that the contrast is the most difficult distortion to handle by the designed classifiers. Although the robustness of both ConvNet and ES-CellNet-1 significantly drops when compared to noise and blur, the ES-CellNet-1 manages to maintain the classification performance at a satisfactory level until class 2 (62.1 % accuracy) whereby the ConvNet only performs somehow only when there is no contrast or only very little distortion (i. e. class 1).

The benchmarking results in all the distortion categories show that ES-CellNet-1 performs better compared to ConvNet plus which is trained to classify MNIST images in the presence of distortions.

3.4 Results: Enhancement of Distorted MNIST Images

The sensitivity curve analysis performed in Section 3.4 justifies the requirement of a preprocessing module (for image pre-enhancement) in order to improve the classification accuracy. To fulfill this requirement, a new neural network, the ES-CellNet-2 enhancement module introduced in the Section 3.3 is utilized. The following Figures 11, 12 and 13 show the image enhancement performance (output image results) of ES-CellNet-2 in the presence of blur, noise and contrast respectively.

The corresponding root mean square error (RMSE) at the testing phase for each distortion is given in the following table (see Table 3).

Table 3: RMSE values corresponding to image enhancement in presence of different distortions at the test phase

Distortion type	RMSE
Gaussian blur	0.0465212773806
Gaussian noise	0.0475116986035
Contrast	0.102221736337

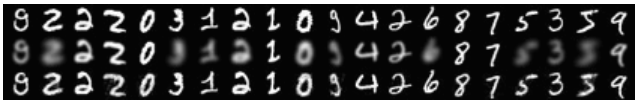


Fig. 11: MNIST enhancement in the presence of gaussian blur. Top row: Target images, Second row: Test images with different levels of Gaussian blur induced and Bottom row: enhanced blur-distorted images

Fig. 12: MNIST enhancement in the presence of gaussian noise. Top row: Target images, Second row: Test images with different levels of Gaussian noise induced and Bottom row: enhanced noise-distorted images

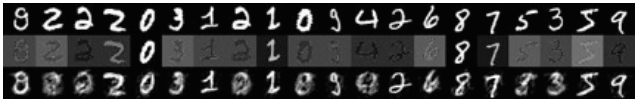


Fig. 13: MNIST enhancement in the presence of contrast. Top row: Target images, Second row: Test images with different levels of contrast induced and Bottom row: enhanced contrast-distorted images

From the results provided in the Figures 11, 12, 13 and the Table 3, it is visible that with the ES-CellNet-2, the enhancement of the images in the presence of blur and noise performs similarly. On the contrary, as shown in the Figure 13, enhancement of the images in the presence of contrast has significant difficulties.

As the next step of the OCR stress testing, the OCRs were tested with the enhanced images. The sensitivity curves of the OCR performance when subjected to enhanced images is as follows (Figure 14, Figure 15 and Figure 16).

The average accuracies of the classifier models when processing an enhanced image are given in the following table 4.

From the Figures 14 and 15, it can be seen that a removal of Gaussian blur and Gaussian noise from the distorted MNIST images improves the OCR accuracy significantly for all the classes. Due to the difficulty of image enhancement in the presence of Contrast (shown in Figure 13 and Table 3), the OCR accuracies

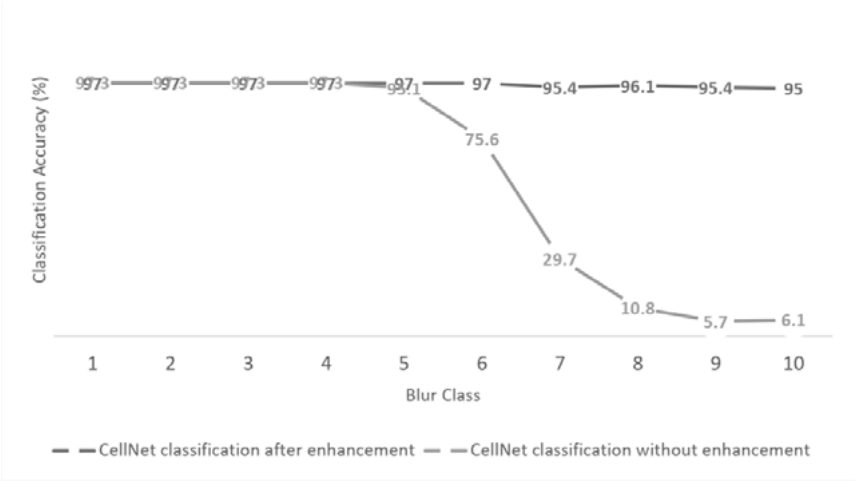


Fig. 14: The blur sensitivity curve showing the difference between accuracies when processing distorted and enhanced images

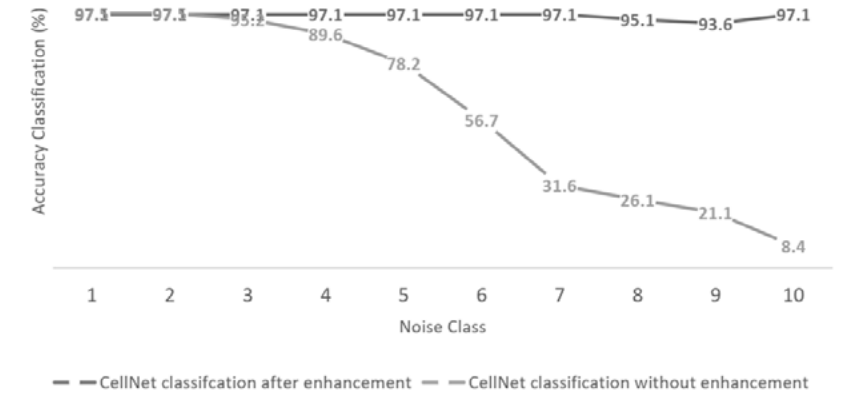


Fig. 15: The noise sensitivity curve showing the difference between accuracies when processing distorted and enhanced images

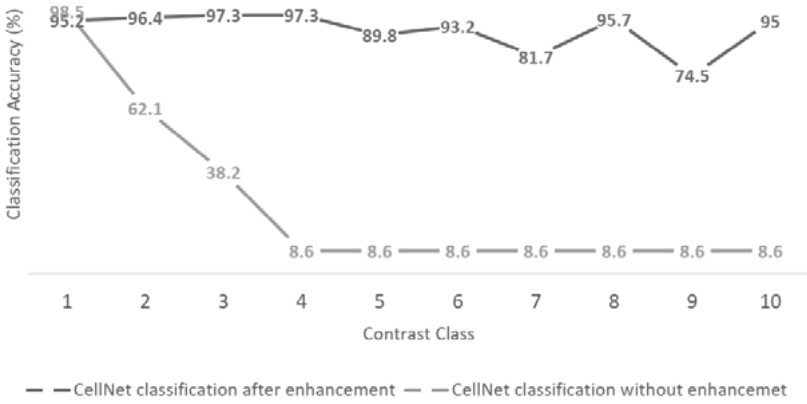


Fig. 16: The contrast sensitivity curve showing the difference between accuracies when processing distorted and enhanced images

Table 4: Average classification accuracies when enhanced images are given to the classifier models (distortions (input images are pre-enhanced)

	Average Accuracy values obtained		
Model	Gaussian Blur	Gaussian Noise	Contrast
ConvNet	99.50 %	98.33 %	96.00 %
ConvNet Plus	99.99 %	94.16 %	99.83 %
ES-CellNet-1	98.00 %	98.00 %	96.55 %

fluctuates even after the images are enhanced (Figure 16). But these accuracy fluctuations remain within an acceptable range proving that even a slight improvement in a distorted image is sufficient for the ES-CellNet-1 to properly perform a character recognition. Furthermore, the results justify that a proper preprocessing system that could enhance the images in the presence of different distortions is necessary for improving the OCR accuracies.

Overall, ConvNet-Plus and ES-CellNet-1 perform much better than ConvNet. In the cases of pre-enhanced images, both ConvNet-Plus and ES-CellNet-1 performs much better with a small superioty of ConvNet-Plus. Nevertheless, ES-CellNet-1 is an architecture with high potential especially due to three facts: (a) lower consumption of computing resources (compared to the very complex

ConvNet structures); (b) easier and faster training; and (c) some additional tuning potential especially if the simple current output layer can be replaced by a 2 or 3 layed MLP neural network structure.

4 Concluding Remarks

This paper first discusses the existence of different types of distortions in a document image and organize them according to the source of occurrence. Then a brief introduction to the OCR system is given in order to provide an insight to the internal dynamics. The mathematical approximation of the identified distortions is presented in this paper and later these approximations are being used to induce the distortions artificially to the MNIST images for dataset preparation.

Using the original MNIST dataset, two classifier models are trained to perform a sensitivity analysis. This analysis shows the variations in the classification accuracies in the presence of different distortions. From the sensitivity analysis it is concluded that the ES-CellNet-1 performs better than the ConvNet. Furthermore, an image enhancer model based on ES-CellNet is designed to justify the requirement of the preprocessing module. In order to gain an understanding on the performance of a ConvNet when trained with distorted images, the ConvNet Plus model is used.

Overall, the classification experiments conducted in this paper concludes that the ES-CellNet-1 and ConvNet Plus perform better than the ConvNet. Regarding the benchmarking results of ES-CellNet-1 against the ConvNet Plus shows that even without training on distorted MNIST images, ES-CellNet-1 can achieve better results in the case of classifying distorted MNIST images. With respect to the enhanced MNIST images, both ConvNet Plus and ES-CellNet-1 performs similarly. Due to its low resource consumption and ease of training ES-CellNet-1 has more potential over ConvNet Plus when it comes to the classification in the presence of image distortions.

The sensitivity analysis of the ES-CellNet-1 when given an enhanced image proves that a preprocessor would significantly improve the OCR accuracy by removing the present distortions from the image.

Acknowledgement

Key results of this paper were obtained in the frame of a project funded by UNIQUARE GmbH, Austria (Project Titel: Dokumenten-OCR-Analyse und Validie-

rung). We thank the UNIQUARE employees Ralf Pichler, Olaf Bouwmeester und Robert Zupan for their precious contributions and support.

References

[1] M. M. Terras, "The Rise of Digitization", *Digitisation Perspectives*, Bd. 46, p. 3, 2011.

[2] A. Bia, R. Muñoz und J. Gómez, "DiCoMo: the digitization cost model", *International Journal on Digital Libraries*, Bd. 11, Nr. 2, p. 141, 2010.

[3] M. R. Carr, "An analysis of the feasibility of a paperless environment – the case of the Mona School of Business", *MANAGEMENT OF ENVIRONMENTAL QUALITY: AN INTERNATIONAL JOURNAL*, Bd. 16, Nr. 4, p. 287, 2005.

[4] M. Black, F. Berard, A. Jepson, W. Newman, E. Saund, G. Socher und M. Taylor, "The digital office: Overview", in: *AAAI Spring Symposium on Intelligent Environments*, Palo Alto, 1998.

[5] Q. N. Vo, S. H. Kim, H. J. Yang und G. Lee, "Binarization of degraded document images based on hierarchical deep supervised network", *Pattern Recognition*, Bd. 74, p. 568, 2018.

[6] P. Sharma und S. Sharma, "Image processing based degraded camera captured document enhancement for improved OCR accuracy", in: *6th International Conference-Cloud System and Big Data Engineering (Confluence)*, 2016.

[7] G. D. F. P. Silva, R. D. Lins, J. M. Silva, S. Banergee, A. Kuchibhotla und M. Thielo, "Enhancing the filtering-out of the back-to-front interference in color documents with a neural classifier", in: *20th International Conference on Pattern Recognition*, 2010.

[8] N. Nayef, M. M. Luqman, S. Prum, S. Eskenazi, J. Chazalon und J.-M. Ogier, "SmartDoc-QA: A Dataset for Quality Assessment of Smartphone Captured Document Images Single and Multiple Distortions", in: *13th International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, 2015.

[9] J. Burie, J. Chazalon, M. Coustaty, S. Eskenazi, M. Luqman, M. Mehri, N. Nayef, J. Ogier, S. Prum und M. Rusinol, "ICDAR2015 competition on smartphone document capture and OCR (SmartDoc)", in: *13th International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, 2015.

[10] F. Asad, A. Ul-Hasan, F. Shafait und A. Dengel, "High Performance OCR for Camera-Captured Blurred Documents with LSTM Networks", in: *12th IAPR Workshop on Document Analysis Systems (DAS)*, Santorini, Greece, 2016.

[11] Q. A. BUI, D. MOLLARD und S. TABBONE, "Selecting automatically preprocessing methods to improve OCR performances", in: *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto, Japan, 2017.

[12] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu und X. Bai, "Multi-oriented text detection with fully convolutional networks", in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, United State of America, 2016.

[13] A. Sulaiman, K. Omar und M. F. Nasrudin, "Degraded Historical Document Binarization: A Review on Issues, Challenges, Techniques, and Future Directions", *Journal of Imaging*, Bd. 5, Nr. 4, pp. 1–7, 2019.

[14] K. Aparna und A. Ramakrishnan, "A complete tamil optical character recognition system", in: *International Workshop on Document Analysis Systems*, Princeton, NJ, United States of America, 2002.

[15] R. Mithe, S. Indalkar und N. Divekar, "Optical character recognition", *International journal of recent technology and engineering (IJRTE)*, Bd. 2, Nr. 1, pp. 73–75, 2013.

[16] M. Ravina, I. Supriya und D. Nilam, "Optical Character Recognition", *International Journal of Recent Technology and Engineering (IJRTE)*, Bd. 2, Nr. 1, p. 74, 2013.

[17] B. Sameeksha, "Optical Character Recognition Using Artificial Neural Network", *International Journal of Advanced Research in Computer Engineering & Technology*, Bd. 1, Nr. 4, p. 131, 2012.

[18] C. H. Strohmaier, C. S. Ringlstetter, U. Klaus und S. Mihov, "Lexical postcorrection of OCR-results: The web as a dynamic secondary dictionary", in: *Seventh International Conference on Document Analysis and Recognition*, 2003. Proceedings, Edinburgh, UK, 2003.

[19] R. Smith, "An overview of the Tesseract OCR engine", in: *Ninth Int. Conference on Document Analysis and Recognition (ICDAR 2007)*, Parana, 2007.

[20] R. Szeliski, "Image processing", in: *Computer Vision: Algorithms and Applications*, Springer, 2010, p. 103.

[21] C.-C. Yang und H. H. Chen, "Gaussian noise approximation for disparity-based autofocus", in: *IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, United States of America, 2016.

[22] P. Singh und R. Shree, "A comparative study to noise models and image restoration techniques", *International Journal of Computer Applications*, Bd. 149, Nr. 1, p. 20, 2016.

[23] H. Hu und G. De Haan, "Low cost robust blur estimator", in: *International Conference on Image Processing*, Atlanta, GA, United States of America, 2006.

[24] P. Singh und R. Shree, "A Comparative Study to Noise Models and Image Restoration Techniques", *International Journal of Computer Applications*, Bd. 149, Nr. 1, p. 19, 2016.

[25] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note", *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, Bd. 148, Nr. 34, p. 5, 2001.

[26] M. Lukosevicius, "A practical guide to applying echo state networks", in: *Neural networks: Tricks of the trade*, Springer, 2012, pp. 2–11.

[27] L. O. Chua und L. Yang, "Cellular neural networks: Theory", IEEE Transactions on circuits and systems, Bd. 35, Nr. 10, pp. 1257–1272, 1988.

[28] G. Grassi, "On discrete-time cellular neural networks for associative memories", *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, Bd. 48, Nr. 1, p. 107, 2001.

On the Introduction of Nonlinear Dynamics in Image Processing with Applications

M. Sarwar Zahan, Jean Chamberlain Chedjou, and Kyandoghere Kyamakya

Institute of Smart Systems Technologies, Transportation Informatics Group
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

Abstract: We apply the concept of nonlinear dynamics to perform specific image processing tasks (e. g., enhancement, edge detection, denoising, conversion to binary image, chaotic encryption, etc.). A class of oscillators is considered each of which is modeled by nonlinear differential equations. Based on these equations the nonlinear dynamical behavior of the oscillators is investigated. This investigation leads to the depiction/discovery of three main states of the oscillators namely, equilibrium, stationary, and chaotic states. The occurrence of these states is predicted and controlled through a bifurcation analysis. Using the bifurcation analysis, ranges/windows of parameters are derived under which each specific oscillator can exhibit the aforementioned states. In each state of the oscillators the image processing is carried out and the output images are obtained as results of processing. Comparing output- with input-images reveals the basic processing tasks performed (e. g., enhancement, edge detection, denoising, conversion to binary image, chaotic encryption, etc.).

1 Introduction

Nowadays, many modern technologies are based on precise image processing algorithms. For example, in traffic management system image processing techniques become very popular because of maintenance simplicity and smart-decision. However, due to the dynamic change of the traffic environment acquiring a good quality image is a challenging task. In most traffic management and driving assistance systems, cameras are used as a major sensing device. If the cameras fail to capture a good quality image, the uncertainties of the system increase significantly. In many cases, the image quality depends on the amount of light available in the dynamic environment. Several factors can be involved in a dynamic environment, for example, foggy environment, the tunnel where light decreases, the movements of large clouds, etc. All these conditions

minimize sunlight and create difficulties for the sensors to capture an effective image. For classical image processing approaches, it is difficult to figure out environmental dynamics to solve the worst image capturing problem. Therefore, the captured images remain with low contrast, blurry, noisy and many more problems. With a poor-quality image, it is difficult for the conventional image processing system to perform complex processing tasks. Moreover improving graphical information for human and autonomous machine perception is the major objective of image processing [1]. Nowadays, many applications depend on high-level image processing, such as automotive and robotics sensing, weather forecasting, medical image analysis, space image analysis, geological data analysis, agricultural crops prediction, etc.. The growing field of image processing faces an enormous amount of challenges, e. g., compression, enhancement, smoothing, etc.. In addition, the processing must meet some critical task requirements. For example, autonomous driving requires real-time data analysis and decision making. Failure to comply with these key requirements may cause fatal circumstances. Similarly, a powerful tracking system also requires a high-speed computing platform to identify the speed of objects and positions. To satisfy this demand, a tremendous attention is nowadays devoted to the development of sensor-based technologies, such as radar sensors, laser sensors, acoustic sensors, etc.. These sensory systems detect the target by calculating travel time of the emitted signal from the sensor and reflected signal from the target [2]. This detection process benefits many target tracking systems. For example, they can easily measure distance, speed or even target locations without additional high-speed computing resources. The disadvantages/drawbacks of these sensory systems are: the very expensive costs, slow scanning speed, high interference and low resolution [3]. In contrast, an optical sensor based-processing system is relatively less costly and can cover around 360-degree views with high resolution. However, dealing with the high-resolution image is a time and resource consuming operation since the pixel per bit ratio increases rapidly. Using a sequential image processing approach to handle large amounts of data is very critical due to limited CPU resources. To handle this bulk amount of data flows, parallel computing approaches become a vital part of image processing [4]. The current generation of multi-core CPUs can process one-megapixel image at high frame rates. However, the CPU acts as a decision engine and performs many tasks simultaneously and this consequently consumes a large amount of CPU resources. Therefore developing efficient image processing algorithms could be an alternative solution to tackle the aforementioned problems encountered when processing images. Indeed an efficient image processing algorithm that can work perfectly in a limited resource environment is highly desirable/recom-

mended. In the last few decades, numerous algorithms have been developed to solve image processing tasks/problems. For example, traditional image processing algorithms (based on statistical pattern recognition techniques) were very popular in the nineties [5]. Nowadays, partial differential equations (PDEs), nonlinear ordinary differential equations (NODEs) and Neural Networks (NNs) become popular tools/instruments for solving image processing problems. The ideas of Neural Networks (i.e., ANN, CNN, and RNN) come from biological neural networks that can easily solve image and word processing problems. Artificial Neural Networks are built in the same way, so they can easily reach biological counterparts [6]. Generally, neural networks are combined with several topologies, such as feed forward, recurrent, continuous-time, discrete-time, synchronous and asynchronous, etc. .

Among the broad/wide family of Neural Networks, Cellular Neural Networks (in their oscillatory functioning mode/state) are gaining tremendous popularity due to their highly efficient computing capabilities. As the name suggests, CNN consists of cells that connect to their neighbors and form a network [7]. Each of these cells has its own dynamics, represented by the state of the cell. This state evolves in time within certain rules, which are generally governed by the strength of coupling between neighboring cells called templates. Due to the coupling between neighboring cells, CNNs computing potential can increase extremely fast, with a computing speed comparable to modern supercomputers [7]. Furthermore, by definition, cells connectivity in CNN processors is local and therefore limited to neighboring cells; this provides simplicity to build VLSI technology. The characteristics of local cell coupling make the CNN architecture very suitable for image processing. The CNN chip behaves like a parallel computer Single-Instruction-Multiple-Data (SIMD). This witnesses the tremendous attention devoted during the past decades to the concept of SIMD as this concept is highly desirable in the field of image processing.

The key focus of this work is to demonstrate how the paradigm of nonlinear oscillators can be applied to image processing. This work presents a certain class of nonlinear and/or chaotic oscillatory systems along with corresponding mathematical models. We present an in-depth demonstration of how the mathematical models can be appropriately/suitably used for image processing. Finally a comparison is performed between the performances of the oscillatory systems at stake. Amongst the oscillators at stake in this work, a novel system [8] is investigated which can overcome many drawbacks/limitations of image processing. Further a focus is devoted to both continuous time and discrete time CNNs. The

processing results show that CNNs can be used efficiently to perform different image processing tasks.

The paper is organized as follow. The second part proposes a description of the methodology consisting of using oscillatory systems for image processing. The third part presents the nonlinear and/or chaotic oscillatory models at stake. A brief description of each oscillator is presented along with the corresponding mathematical model. The numerical simulation of the mathematical models describing the oscillators is carried out leading to the depiction of the processed image (as result of the image processing task performed). The fifth part is devoted to concluding remarks.

2 Methodology: Image Processing based on Oscillatory Theory

The flow diagram summarizing all key steps of the methodology of using nonlinear and/or chaotic oscillators for image processing is depicted in Fig. 1. The image processing-based nonlinear oscillators is straightforward and the processing steps remain unchanged for all types of oscillators. The key steps are summarized into six points as shown in Fig. 1.

3 Oscillators at Stake and Results of Numerical Processing

3.1 Chaotic Encryption based on the Lorenz Oscillator

The Lorenz system was proposed by Edward Lorenz in 1963 [9] in the form of coupled ordinary differential equations (see (1)). The coupled system (1) is extremely sensitive to initial conditions. Further for certain parameter values and initial conditions, the system may undergo chaotic states. During the discovery of the system in (1) Edward Lorenz was interested in modeling the complex behavior of the Earth's atmosphere with many parameters. He simplified the model with a single atmospheric condition called rolling fluid convection. According to him, convections can occur from large to small scale. Always considered by the scientific community as "father" or "founder" of chaos theory, the theory was shortly defined by Edward Lorenz as recalled in [10] – "Chaos: When the present determines the future, but the approximate present does not approximately determine the future". Based on this theory the scientific community has devoted a tremendous attention to the investigation of the chaotic

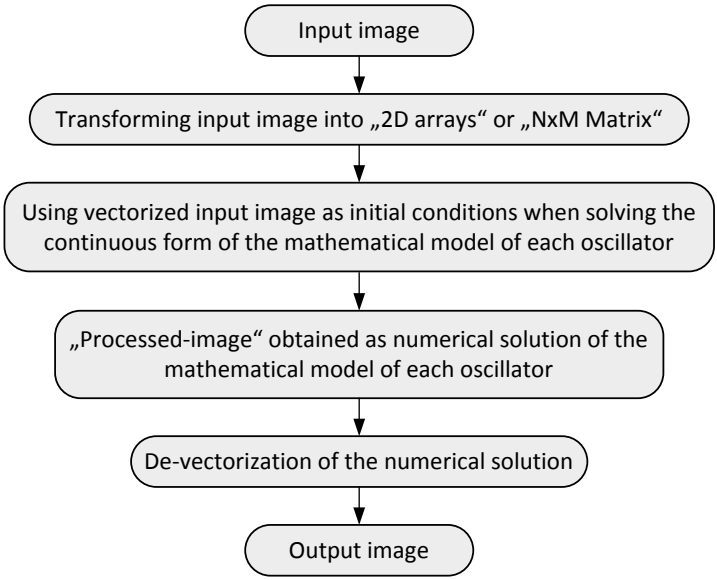


Fig. 1: Flow diagram of the key steps involved in the image processing-based oscillatory theory

behavior of many natural systems such as weather condition, road traffic, etc. .

$$\begin{aligned} \frac{dx}{dt} &= \sigma(y-x) \\ \frac{dy}{dt} &= x(\rho-z)-y \\ \frac{dz}{dt} &= xy-\beta z \end{aligned} \tag{1}$$

For a specific case, σ is the ratio between thermal conductivity, ρ is the temperature difference and β is the horizontal diameter of the roll. Furthermore, x, y, and z are the convective flow that evolves in time with respect to the horizontal and vertical temperature distributions.

The phase portrait (x, z) in Fig. 2 is obtained as direct numerical simulation of (1) using the parameter settings $\alpha = 10$, $\rho = 28$ and $\beta = 8/3$. This phase portrait witnesses the well-known shape of the chaotic Lorenz oscillator. In the chaotic state of the Lorenz oscillator, the system (1) is used for image processing

according to the flow diagram in Fig. 1. The results obtained correspond to the output image in Fig. 2. The histogram of the output image is presented in Fig. 2. This histogram confirms the chaotic state of the oscillator. Therefore the output image is a chaotic encryption of the input (original) image.

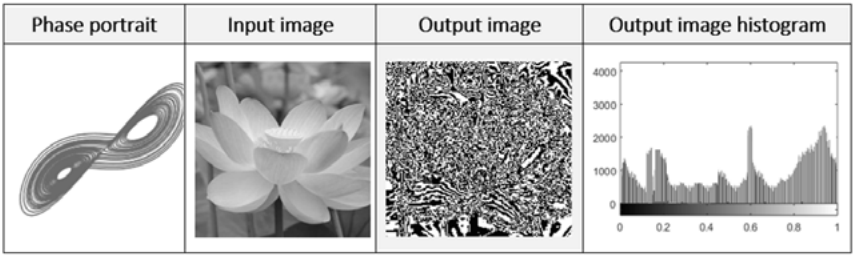


Fig. 2: Result of image processing through Lorenz oscillator. The parameter settings are $\alpha = 10$, $\rho = 28$ and $\beta = 8/3$. Numerical phase portrait (x, z) and input image are in first and second columns. Output image and corresponding histogram are in third and fourth columns.

3.2 Chaotic Encryption based on a Coupling of the van der Pol and the Duffing Oscillators

In recent decades, several nonlinear coupled oscillators have been discovered to meet the demands of science and engineering such as adaptive frequency oscillator (Hopf), van der Pol oscillator coupled to Duffing oscillator, etc.. The methods of nonlinear coupled oscillatory systems has been devoted a great attention, especially in the field of image processing. However, very few articles have been published on non-linear coupled oscillators that provide detailed information on image processing. The system consisting of the van der Pol oscillator coupled to Duffing oscillator discovered by Jean Chamberlain Chedjou [8] shows a very promising approach that can solve contrast enhancement problem in dynamic environmental conditions. Indeed the system is adaptive to dynamic variations in ambient light. The work in [12] also reveals that the proposed oscillatory model is extremely robust especially when it comes to real time scenarios. Most oscillatory image processing techniques are designed for static images that limit the application area. In contrast, the van der Pol oscillator coupled to Duffing oscillator is a novel approach to image processing that can be applied to dynamic images. The oscillatory model of the literature [8] keeps

the parameter of dissipative coupling c_2 and c_4 equal zero for the local stability analysis (i. e., analysis of equilibrium/fixed points). In this example, the van der Pol oscillator coupled to Duffing oscillatory system parameter setting are taken from [12] for further analysis of the system behavior with application in image processing. The mathematical model of this oscillatory paradigm is expressed through the following coupled equations:

$$\begin{aligned} \frac{d^2x}{dt^2} - \epsilon_1(1-x^2)\frac{dx}{dt} + \omega_1^2 x &= c_1 y + c_2\frac{dy}{dt} \\ \frac{d^2y}{dt^2} + \epsilon_2\frac{dy}{dt} - \omega_2^2 y + c_0 y^3 &= c_3 x + c_4\frac{dx}{dt} \end{aligned} \tag{2}$$

The variables and parameters in (2) are defined as follows: x is the solution of the van der Pol oscillator, y is the solution of the Duffing oscillator, ϵ_1 and ϵ_2 are dissipative coefficients, ω_1 is the natural frequency of van der Pol oscillator, ω_2 is the natural frequency of Duffing oscillator, c_1 and c_3 are elastic couplings, c_2 and c_4 are dissipative coupling. The direct numerical simulation of (2) is performed using the following values of parameters: $\epsilon_1 = 0.74431300$, $\epsilon_2 = 0.66808300$, $c_0 = 0.22237500$, $c_1 = 0.000000$, $c_2 = 0.23519100$, $c_3 = 0.000000$, $c_4 = 0.981204$, $\omega_1 = 0.99161434$ and $\omega_2 = 1.264911$. The phase portrait (x, y) in Fig. 3 is obtained as direct numerical simulation of (2). This phase portrait witnesses the well-known shape of the coupled chaotic system as reported in [12]. The chaotic state of the coupled system (2) is used for image processing according to the flow diagram in Fig. 1. The results obtained correspond to the output image in Fig. 3. The histogram of the output image is presented in Fig. 3. This histogram confirms the chaotic state of the oscillator. Therefore the output image is a chaotic encryption of the input (original) image. For the image processing performed in this example, we have used a RGB image 225x315 (pixel) as input source. The initial simulation time used is $t_i = 0$, the step size (or increment time) is $dt = 0.1$ and the simulation interval (duration) is $t_f = 0.25$.

3.3 Denoising based on Continuous Time Cellular Neural Networks (CT-CNN)

Most Cellular Neural Networks (CNN) algorithms are designed to handle specific image processing tasks. Due to the improved parallel processing capabilities and deployment flexibility, CNN has become extremely popular in the field of image and video processing. In convolution form, the mathematical model of the CT-CNN can be written as follows:

$$\dot{x}_{ij} = -x_{ij} + T_C \otimes x_{ij} + T_A \otimes y_{ij} + T_B \otimes u_{ij} + I_{ij} \tag{3}$$

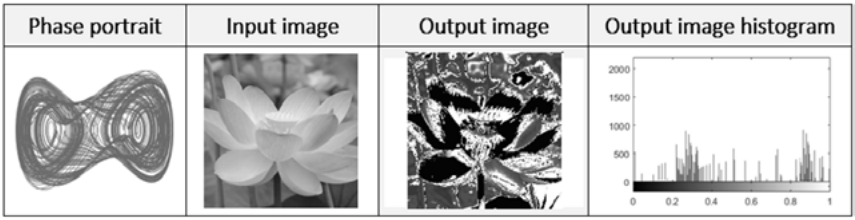


Fig. 3: Result of image processing through a system consisting of the van der Pol oscillator coupled to Duffing oscillator. The parameter settings are defied in the text. Numerical phase portrait (x, y) and input image are in first and second columns. Output image and corresponding histogram are in third and fourth columns.

Where x_{ij} is the state of cell in position (i, j), $\dot{x}_{ij}$ is the time derivate of the state of the cell in position (i, j), y_{ij} represents the nonlinear sigmoid function, i. e. $y_{ij} = (|x+1| - |x-1|)/2$, T_A is the output template, T_B is the input template, T_C is the state-controlled template, u_{ij} is the input of cell in position (i, j), I_{ij} is the bias of cell in position (i, j).

Let's consider $u(x, y)$ is an RGB image (i. e., the input image to be denoised). According to the CNN convention, the grayscale image is ranging from $[-1, +1]$. The image is represented as a $M \times N$ matrix u in the sampled space. The grayscale image of CNN is donated by $u \in [-1, +1]^{M \times N}$ where $M \times N$ is the size of the image. The following CNN templates are used in the current example to perform the image denoising task.

$$T_A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad T_B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad T_C = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -0.1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad I = -1.75$$

Further the following settings are taken into account during the numerical simulation (i. e. image processing): Image size: 225x315; Simulation start time: 0; Simulation increment time: 0.1; Simulation convergence time: 5. The simulation is performed through the MATLAB (2015b) software environment. The Global processing steps of contrast enhancement based on the CT-CNN paradigm are described as follows:

- Load a static RGB image and convert to grayscale for the convolution operation.
- Convert gray scale image into double type for high precision.

- Perform a convolution operation on (double-type) image using input template T_B.
- Convert double-type image into vector image.
- Set the vector image in the initial condition of CNN.
- Consider the state of CNN (expressed in the form of vector image): reshape the vector image into matrix and perform convolution operation using the templates (T_A, T_C). In the case of the input template T_A, convolution operation is performed through a nonlinear function called sigmoid. The solution is vectorized again for the further processing steps.
- The output is obtained in terms of the state through the nonlinear activation function (Sigmoid).
- Finally, the solution expressed in vector form is transformed into a matrix and this transformation produces the desired output (as result of the processing; this corresponds to the processed image based on CT-CNN). The result obtained through the processing based on CT-CNN is clearly illustrated in figure 4.

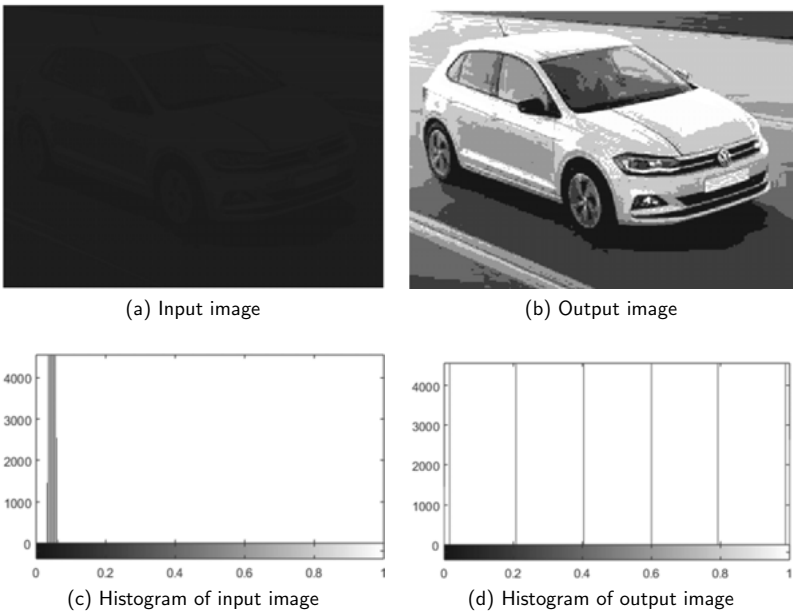
The results in Fig. 4 clearly witness the efficiency of the continuous time cellular Neural Networks (CT-CNN) for image denoising. Note that the denoising here strongly depends on the choice of the parameter settings of (3) called templates. Therefore the dynamics of the CT-CNN depends essentially on the values assigned to templates. These templates express the strength of coupling between cells (i. e., direct coupling between neighboring cells and indirect coupling otherwise).

3.4 Contrast Enhancement based on Discrete Time Cellular Neural Networks (DT-CNN)

The mathematical model of the DT-CNN is obtained through discretization of the mathematical model of the CT-CNN in (3). The discretization of (3) leads to the discrete form (4). The expression (4) represents the mathematical model of the DT-CNN.

$$x_n = (x_0 - A - B - C - I)\,(1-h)^n + A + B + C + I \tag{4}$$

In (4), x_n is the solution at the n^{th} discrete point. A is denoted by $T_A \otimes y_{ij}$, B is denoted by $T_B \otimes u_{ij}$, C is denoted by $T_C \otimes x_{ij}$, h is the step size, x_0 is the



(a) Input image (b) Output image

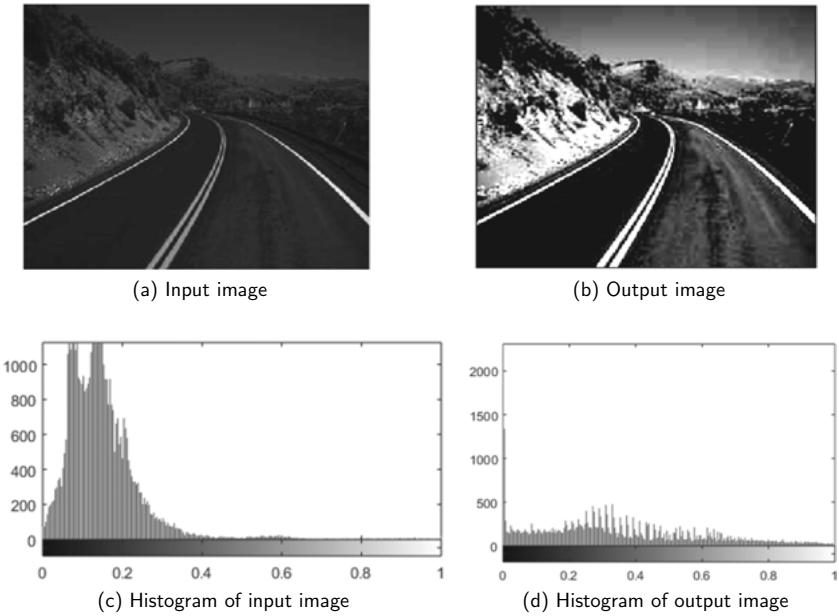
(c) Histogram of input image (d) Histogram of output image

Fig. 4: Result of CT-CNN based image denoising.

original image (used as initial condition at first iteration), and I is a real number representing the threshold.

In this example, the image size is (201x260). In addition, the initial step $n = 1$ is increased by $n + 1$, the total number of iterations (i. e., the maximum value of n) is 60 and the step size is $h = 0.005$. The simulation is performed through MATLAB (2015b) and the template values defined in the case of CT-CNN remain unchanged. Using the values of parameters defined in the case of CT-CNN, the image processing has led to results in Fig. 5. Figure 5a shows the input image of DT-CNN and figure 5b is the corresponding output image. The output image in figure 5b shows that the image visibility has been improved significantly. Therefore, the white lines and side elements of the road become more visible. As illustrative example, the snow on the road in the output image (see figure 5b) is easily identifiable. Figure 5c is the histogram of the input image. This histogram shows that most pixels are distributed/spread between the gray level 0 and 0.4. Therefore, the image is poorly visible. After the processing through DT-CNN, the image pixels are scattered between the gray level 0 and 1. This result is de-

picted in figure 5d. In addition, figure 5d) shows that most of the pixels fall over 0 intensity level. On the other hand, only a few pixels are distributed between 0.6 and 1; that improves visibility of the output image.



(a) Input image (b) Output image

(c) Histogram of input image (d) Histogram of output image

Fig. 5: Result of image contrast enhancement through DT-CNN.

3.5 Edge Detection based on Discrete Time Cellular Neural Networks (DT-CNN)

The mathematical model of the DT-CNN is used in this application example to perform edge detection. The following settings of parameters are considered for the numerical simulation: the image size is (185x186); the maximum number of iterations is $n = 151$; the step size is $h = 0.005$. The simulation is performed through the MATLAB (2015b) software environment. The summary of the steps of the DT-CNN based edge detection process is as follows:

- Load a static grayscale image as system input.
- Convert grayscale image into double type for more precision.

- Perform a convolution operation without any nonlinear function applied to the input template T_B, the output templates T_A and the auxiliary templates T_C.

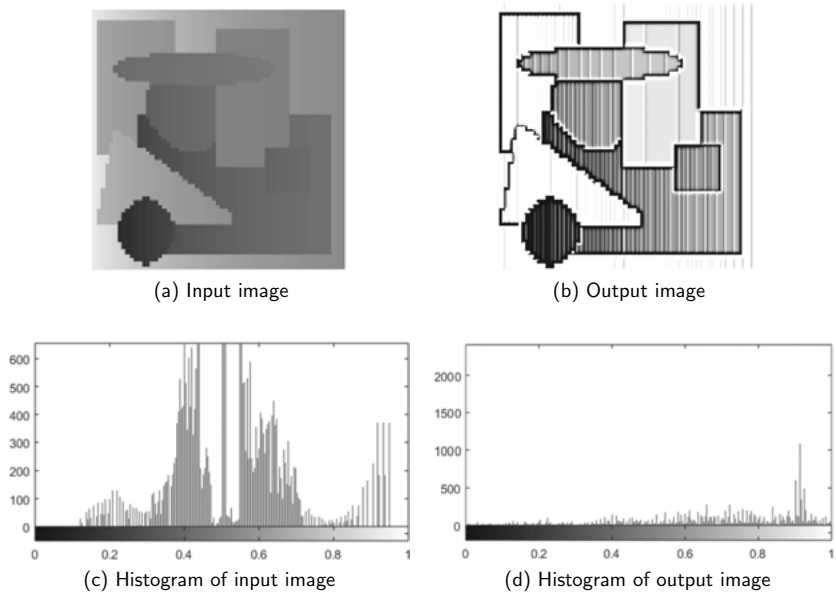
- Resize the image with original image size.

- Finally, reshape the vector to matrix image for desired output.

Figure 6a shows a grayscale input image which is inserted into the DT-CNN model for the edge detection process. The model uses the values of templates define in the text (see above). The corresponding figure 6b represents the output image processed through the DT-CNN model. The output image in figure 6b provides information about the detected edges that help to recognize individual features of the image. The histogram of the input image shows that most pixels are distributed between the gray level 0.3 and 0.7. To detect the edges, discontinuities at the intensity level need to be recognized/identified. Figure 6c is the histogram of the input image and the histogram of the output image is depicted in figure 6d. This later histogram shows that only few pixels are spread/distributed between the gray level 0 and 1. Here, the edges are clearly highlighted and unnecessary pixels are removed. This witnesses the efficiency of the DT-CNN for edge detection.

4 Conclusion

In this work, we have considered a class of nonlinear oscillators, which have been further successfully applied to image processing. Different types of nonlinear oscillators have been considered along with their respective mathematical models. Based on the numerical simulation of the mathematical models (at stake in this work), several concrete tasks/applications have been identified depending on the type of oscillator used. It has been demonstrated that the oscillators at stake efficiently perform the tasks of image encryption, denoising, enhancement, and edge detection, just to name a few. Regarding the CNN based processing, we have presented two paradigms of CNN, namely CT-CNN and DT-CNN. These two CNN paradigms have been used to perform image enhancement, denoising, and edge detection. Simulation results clearly indicate that the global dynamics of the CNN computing platform as well as the performances of CNN are governed by the coupling coefficients called CNN-templates. This clearly witnesses the fact that CNN templates have a significant effect on



(a) Input image (b) Output image

(c) Histogram of input image (d) Histogram of output image

Fig. 6: Result of edge detection through DT-CNN.

the desired outcome of given image processing task. Although the template development is still an active area of research, it has been clearly demonstrated that image processing with three templates provides appealing results.

References

[1] F. S. Amit Kumar, in: Image Processing in Diabetic Related Causes, 1 ed., Springer Singapore, 2016, pp. VI, 56.

[2] Z. Sun, G. Bebis, R. Miller, "On-road vehicle detection: a review", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28 (5):694–711, 2006.

[3] R. Brown, "Generalizations of the Chua equations", IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, Vol. 40 (11):878–884, 1993.

[4] B. Reddy, "Performance Analysis of GPU V/S CPU for Image Processing Applications", International Journal for Research in Applied Science and Engineering Technology, pp. 437–443, 2017.
[5] A. Joshi, N. Ramakrishman, E. Houstis and J. and Rice, "On neurobiological, neuro-fuzzy, machine learning, and statistical pattern recognition techniques", IEEE Transactions on Neural Networks, vol. 8, no. 1, pp. 18–31, 1997.
[6] M. Hänggi, "Analysis, design, and optimization of cellular neural networks", Diss. ETH Zurich, 1999.
[7] L. Chua and T. Roska, Cellular neural networks and visual computing: foundations and applications, Cambridge University Press, June 10, 2002.
[8] Kyamakya, K., J.C. Chedjou and M.A. Latif, and U.A. K, "Cellular Nanoscale Networks and Their Applications (CNNA),12th International Workshop", Berkeley, California, 2010.
[9] Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Lorenz_system. [Accessed 23 july 2018].
[10] C. M. Danforth, "Chaos in an Atmosphere Hanging on a Wall", Mathematics of Planet Earth, 2013.
[11] B. M. I. Haque, "A New Approach of Iteration Method for Solving Some", vol. 13, no. 11, 2013.
[12] J.C. Chedjou, H.B. Fotsin, P. Woafo, S. Domngang, "Analog Simulation of the Dynamics of a van der Pol Oscillator Coupled to a Duffing Oscillator", IEEE Transactions on Circuits and Systems I, vol. 48, issue 6, pp. 748–757, 2001.

Index of Authors

Die Reihen der Fortschritt-Berichte VDI:

1 Konstruktionstechnik/Maschinenelemente
2 Fertigungstechnik
3 Verfahrenstechnik
4 Bauingenieurwesen
5 Grund- und Werkstoffe/Kunststoffe
6 Energietechnik
7 Strömungstechnik
8 Mess-, Steuerungs- und Regelungstechnik
9 Elektronik/Mikro- und Nanotechnik
10 Informatik/Kommunikation
11 Schwingungstechnik
12 Verkehrstechnik/Fahrzeugtechnik
13 Fördertechnik/Logistik
14 Landtechnik/Lebensmitteltechnik
15 Umwelttechnik
16 Technik und Wirtschaft
17 Biotechnik/Medizintechnik
18 Mechanik/Bruchmechanik
19 Wärmetechnik/Kältetechnik
20 Rechnerunterstützte Verfahren (CAD, CAM, CAE CAQ, CIM ...)
21 Elektrotechnik
22 Mensch-Maschine-Systeme
23 Technische Gebäudeausrüstung

ISBN 978-3-18-3**86410**-2