DAS 2022

15th IAPR International Workshop on Document Analysis Systems

May 22-25, La Rochelle, France

The Winner Takes It All
choosing the “best” binarization algorithm for photographed documents

Rafael Dueire Lins, Rodrigo Barros Bernardino,
Ricardo Barboza and Raimundo Oliveira

BRAZIL

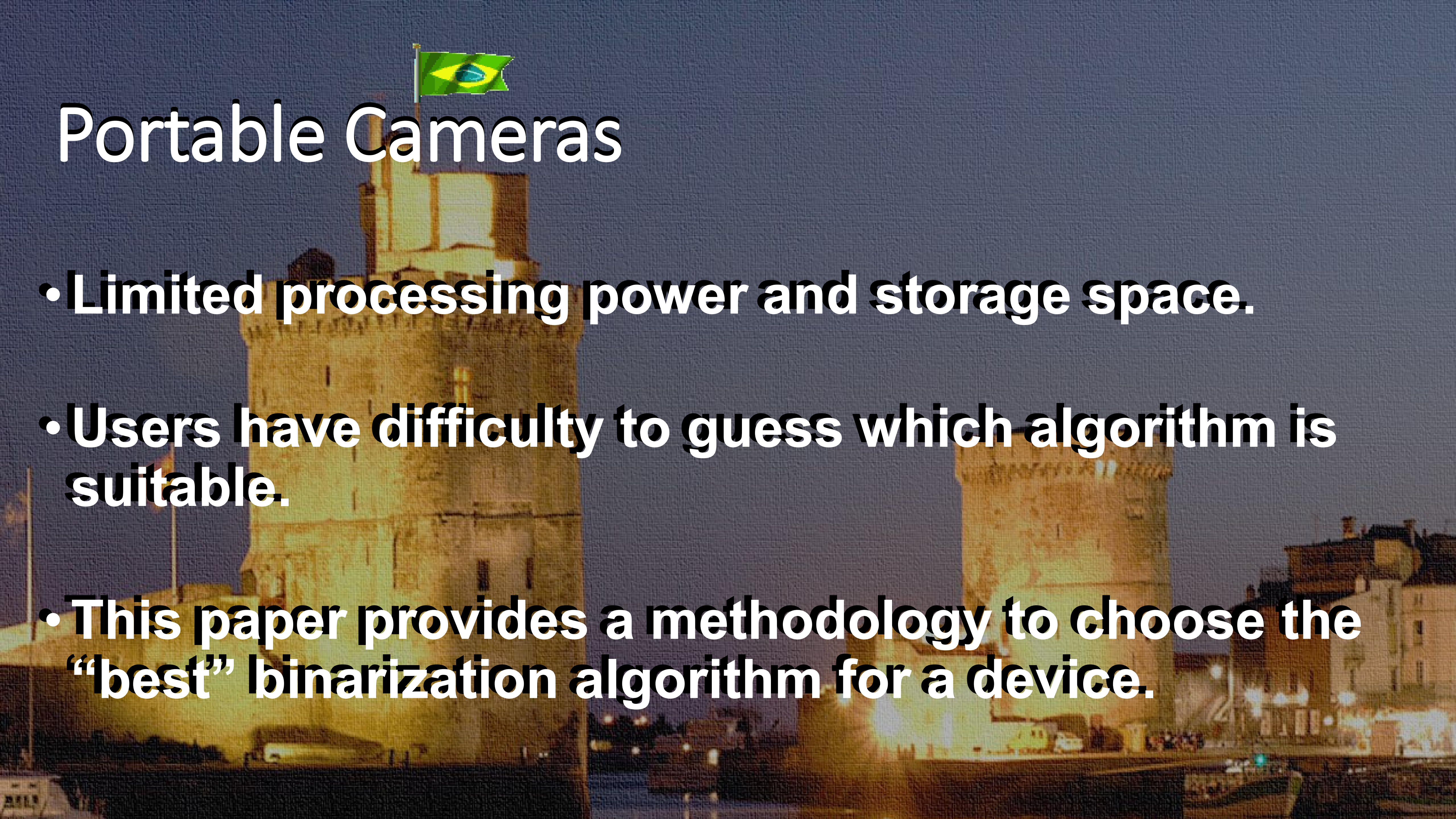
Document Binarization

- Conversion of a color image into black-and-white.
- Makes most documents more readable
- Saves toner for printing.
- Saves storage space.
- Saves communication bandwidth.
- Is a key preprocessing step for document OCR, classification and indexing

Time-Quality Binarization Competitions

- No algorithm is good for all kinds of documents.
- The quality of the resulting image depends on the features of each image.
- Time performance is important for applicability!

Portable Cameras



- Limited processing power and storage space.
- Users have difficulty to guess which algorithm is suitable.
- This paper provides a methodology to choose the "best" binarization algorithm for a device.

Photographed Documents

- **Uneven resolution.**
- **Perspective distortion.**
- **Non-uniform document illumination.**
- **External interfering light sources.**
- **Undesirable non-uniform document framing.**
- **Default file-format: JPEG 1% loss**

61 Algorithms Assessed:

Akbari1-3, Bataineh, Bernsen, Bradley, Calvo-Z, CLD, CNW, dSLR, DeepOtsu, DiegoPavan, DilatedUNet, DocDLink, Doc-UNet, ElisaTV, ErginaG, ErginaL, Gattal, Gosh, Howe, Huang, HuangBCD, HuangUnet, iNICK, Intermodes, ISauvola, IsoData, Jia-Shi, Johannsen, KSW, Li-Tam, Lu-Su, Mean, Mello-Lins, Michalak, Michalak211-213, MinError, Moments, Niblack, Nick, Otsu, Percentile, Pun, RenyEntropy, Sauvola, Shanbhag, Singh, Su-Lu, Triangle, Vahid, WAN, Wolf, Wu-Lu, Yen-CC, YinYang,YinYang21, Yuleny.

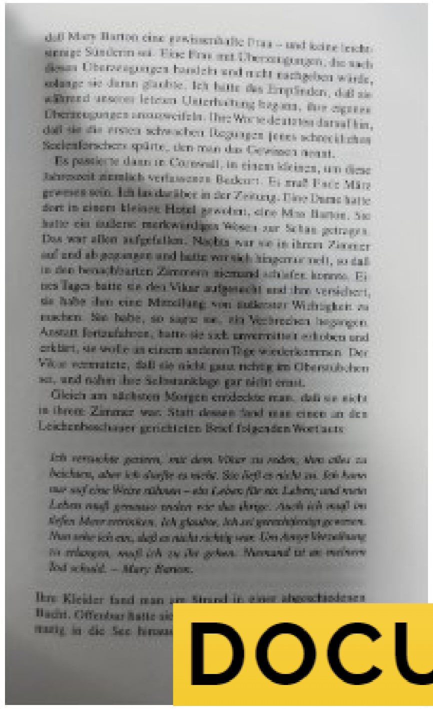
Test Images

Devices:
Motorola G9 Plus,
iPhone SE 2,
Samsung A10S,
Samsung S20

Strobe flash:
top "off",
bottom "on"

TC-10/TC-11
Dataset

IAPR
International Association
for Pattern Recognition

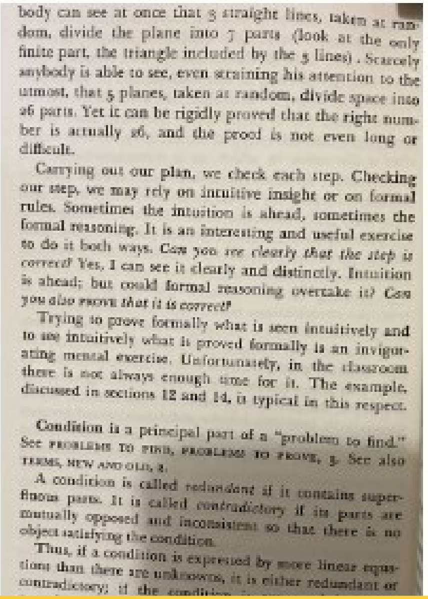

daß Mary Barton eine gewissenhafte Frau – und keine leichtsinnige Sünderin sei. Eine Frau mit Überzeugungen, die nach diesen Überzeugungen handeln und nicht nachgeben würde, solange sie daran glaubte. Ich hatte das Empfinden, daß sie während unserer letzten Unterhaltung begann, ihre eigenen Überzeugungen anzuzweifeln. Ihre Worte deuteten darauf hin, daß sie die ersten schwachen Regungen jenes schrecklichen Seelenforschers spürte, den man das Gewissen nennt.

Es passierte dann in Cornwall, in einem kleinen, um diese Jahreszeit ziemlich verlassenen Badeort. Es muß Ende März gewesen sein. Ich las darüber in der Zeitung. Eine Dame hatte dort in einem kleinen Hotel gewohnt, eine Miss Barton. Sie hatte ein äußerst merkwürdiges Wesen zur Schau getragen. Das war allen aufgefallen. Nachts war sie in ihrem Zimmer auf und ab gegangen und hatte vor sich hingemurmelt, so daß in den benachbarten Zimmern niemand schlafen konnte. Eines Tages hatte sie den Vikar aufgesucht und ihm versichert, sie habe ihm eine Mitteilung von äußerster Wichtigkeit zu machen. Sie habe, so sagte sie, ein Verbrechen begangen. Anstatt fortzufahren, hatte sie sich unvermittelt erhoben und erklärt, sie wolle an einem anderen Tage wiederkommen. Der Vikar vermutete, daß sie nicht ganz richtig im Oberstübchen sei, und nahm ihre Selbstanklage gar nicht ernst.

Gleich am nächsten Morgen entdeckte man, daß sie nicht in ihrem Zimmer war. Statt dessen fand man einen an den Leichenbeschauer gerichteten Brief folgenden Wortlauts:

Ich versuchte gestern, mit dem Vikar zu reden, ihm alles zu beichten, aber ich durfte es nicht. Sie ließ es nicht zu. Ich kann nur auf eine Weise sühnen – ein Leben für ein Leben; und mein Leben muß genauso enden wie das ihrige. Auch ich muß im tiefen Meer ertrinken. Ich glaubte, ich sei gerechtfertigt gewesen. Nun sehe ich ein, daß es nicht richtig war. Um Annys Verzeihung zu erlangen, muß ich zu ihr gehen. Niemand ist an meinem Tod schuld. – Mary Barton.

Ihre Kleider fand man am Strand in einer abgeschiedenen Bucht. Offenbar hatte sie mutig in die See hinaus


body can see at once that 3 straight lines, taken at random, divide the plane into 7 parts (look at the only finite part, the triangle included by the 3 lines). Scarcely anybody is able to see, even straining his attention to the utmost, that 5 planes, taken at random, divide space into 26 parts. Yet it can be rigidly proved that the right number is actually 26, and the proof is not even long or difficult.

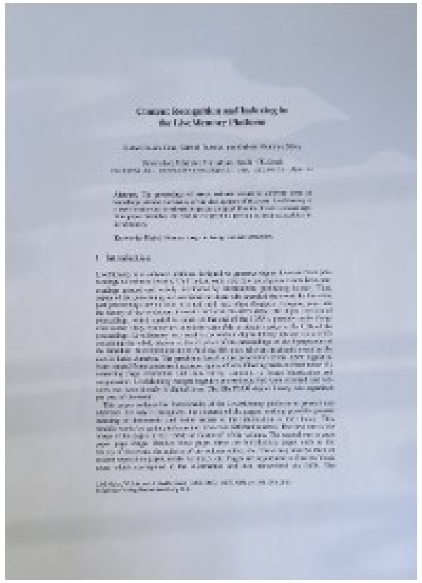
Carrying out our plan, we check each step. Checking our step, we may rely on intuitive insight or on formal rules. Sometimes the intuition is ahead, sometimes the formal reasoning. It is an interesting and useful exercise to do it both ways. *Can you see clearly that the step is correct?* Yes, I can see it clearly and distinctly. Intuition is ahead; but could formal reasoning overtake it? *Can you also* PROVE *that it is correct?*

Trying to prove formally what is seen intuitively and to see intuitively what is proved formally is an invigorating mental exercise. Unfortunately, in the classroom there is not always enough time for it. The example, discussed in sections 12 and 14, is typical in this respect.

Condition is a principal part of a "problem to find." See PROBLEMS TO FIND, PROBLEMS TO PROVE, 3. See also TERMS, NEW AND OLD, 2.

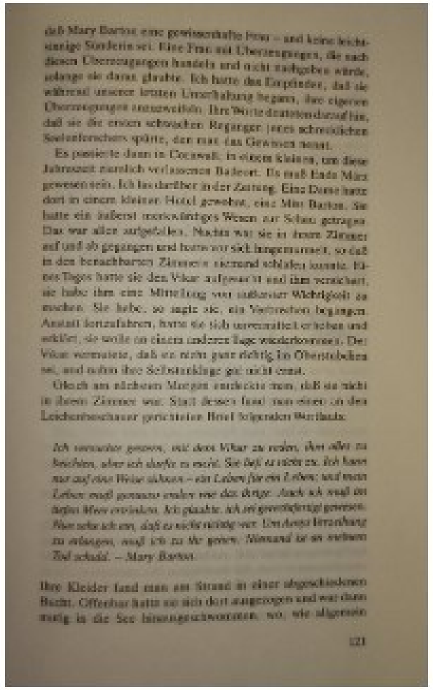
A condition is called *redundant* if it contains superfluous parts. It is called *contradictory* if its parts are mutually opposed and inconsistent so that there is no object satisfying the condition.

Thus, if a condition is expressed by more linear equations than there are unknowns, it is either redundant or contradictory; if the condition



DOCUMENT IMAGE BINARIZATION

https://dib.cin.ufpe.br/



daß Mary Barton eine gewissenhafte Frau – und keine leichtsinnige Sünderin sei. Eine Frau mit Überzeugungen, die nach diesen Überzeugungen handeln und nicht nachgeben würde, solange sie daran glaubte. Ich hatte das Empfinden, daß sie während unserer letzten Unterhaltung begann, ihre eigenen Überzeugungen anzuzweifeln. Ihre Worte deuteten darauf hin, daß sie die ersten schwachen Regungen jenes schrecklichen Seelenforschers spürte, den man das Gewissen nennt.

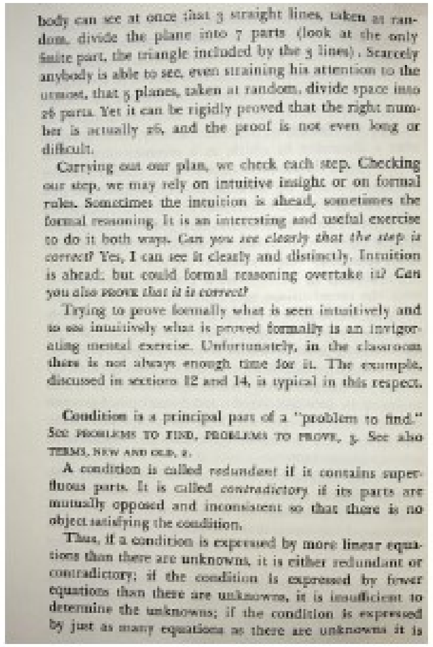
Es passierte dann in Cornwall, in einem kleinen, um diese Jahreszeit ziemlich verlassenen Badeort. Es muß Ende März gewesen sein. Ich las darüber in der Zeitung. Eine Dame hatte dort in einem kleinen Hotel gewohnt, eine Miss Barton. Sie hatte ein äußerst merkwürdiges Wesen zur Schau getragen. Das war allen aufgefallen. Nachts war sie in ihrem Zimmer auf und ab gegangen und hatte vor sich hingemurmelt, so daß in den benachbarten Zimmern niemand schlafen konnte. Eines Tages hatte sie den Vikar aufgesucht und ihm versichert, sie habe ihm eine Mitteilung von äußerster Wichtigkeit zu machen. Sie habe, so sagte sie, ein Verbrechen begangen. Anstatt fortzufahren, hatte sie sich unvermittelt erhoben und erklärt, sie wolle an einem anderen Tage wiederkommen. Der Vikar vermutete, daß sie nicht ganz richtig im Oberstübchen sei, und nahm ihre Selbstanklage gar nicht ernst.

Gleich am nächsten Morgen entdeckte man, daß sie nicht in ihrem Zimmer war. Statt dessen fand man einen an den Leichenbeschauer gerichteten Brief folgenden Wortlauts:

Ich versuchte gestern, mit dem Vikar zu reden, ihm alles zu beichten, aber ich durfte es nicht. Sie ließ es nicht zu. Ich kann nur auf eine Weise sühnen – ein Leben für ein Leben; und mein Leben muß genauso enden wie das ihrige. Auch ich muß im tiefen Meer ertrinken. Ich glaubte, ich sei gerechtfertigt gewesen. Nun sehe ich ein, daß es nicht richtig war. Um Annys Verzeihung zu erlangen, muß ich zu ihr gehen. Niemand ist an meinem Tod schuld. – Mary Barton.

Ihre Kleider fand man am Strand in einer abgeschiedenen Bucht. Offenbar hatte sie sich dort ausgezogen und war dann mutig in die See hinausgeschwommen, so, wie allgemein

121



body can see at once that 3 straight lines, taken at random, divide the plane into 7 parts (look at the only finite part, the triangle included by the 3 lines). Scarcely anybody is able to see, even straining his attention to the utmost, that 5 planes, taken at random, divide space into 26 parts. Yet it can be rigidly proved that the right number is actually 26, and the proof is not even long or difficult.

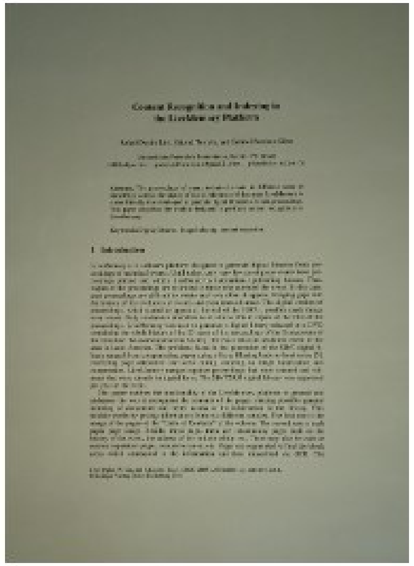
Carrying out our plan, we check each step. Checking our step, we may rely on intuitive insight or on formal rules. Sometimes the intuition is ahead, sometimes the formal reasoning. It is an interesting and useful exercise to do it both ways. *Can you see clearly that the step is correct?* Yes, I can see it clearly and distinctly. Intuition is ahead; but could formal reasoning overtake it? *Can you also* PROVE *that it is correct?*

Trying to prove formally what is seen intuitively and to see intuitively what is proved formally is an invigorating mental exercise. Unfortunately, in the classroom there is not always enough time for it. The example, discussed in sections 12 and 14, is typical in this respect.

Condition is a principal part of a "problem to find." See PROBLEMS TO FIND, PROBLEMS TO PROVE, 3. See also TERMS, NEW AND OLD, 2.

A condition is called *redundant* if it contains superfluous parts. It is called *contradictory* if its parts are mutually opposed and inconsistent so that there is no object satisfying the condition.

Thus, if a condition is expressed by more linear equations than there are unknowns, it is either redundant or contradictory; if the condition is expressed by fewer equations than there are unknowns, it is insufficient to determine the unknowns; if the condition is expressed by just as many equations as there are unknowns it is



Smartphones:

- **Mid-price range models of different manufacturers.**
- **Used by millions of people!**

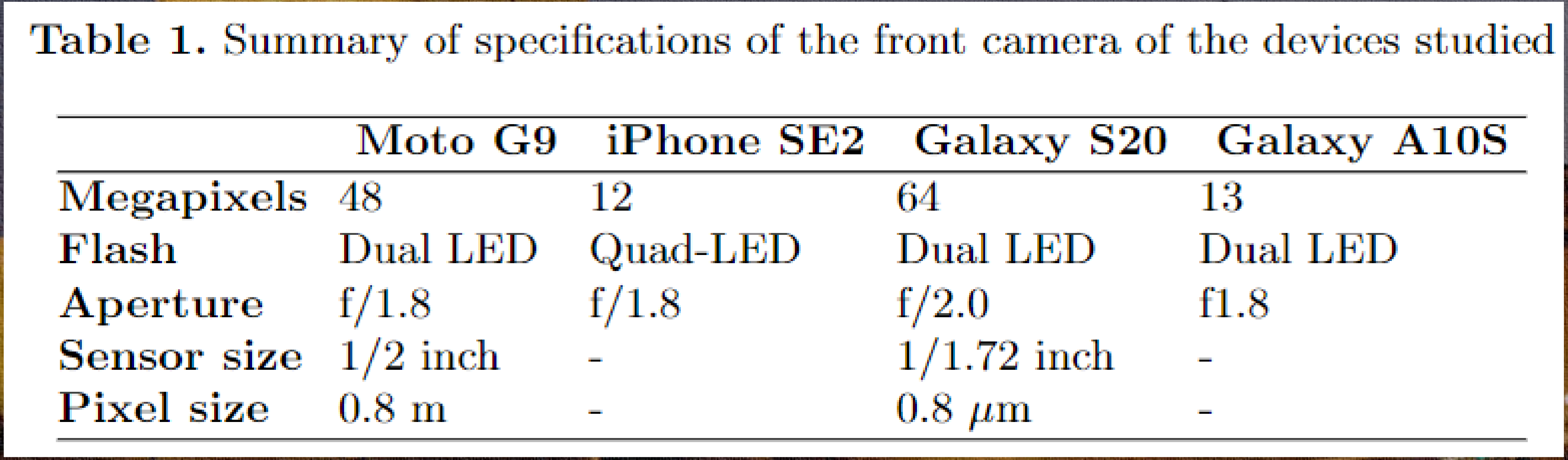
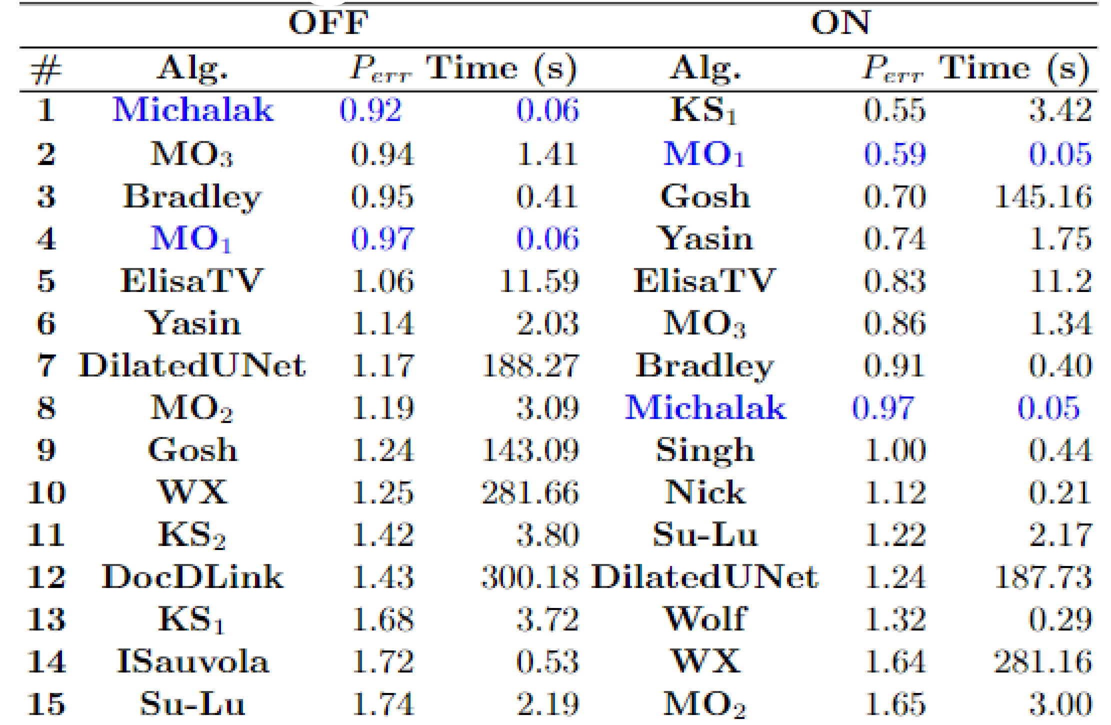


Table 1. Summary of specifications of the front camera of the devices studied

	Moto G9	iPhone SE2	Galaxy S20	Galaxy A10S
Megapixels	48	12	64	13
Flash	Dual LED	Quad-LED	Dual LED	Dual LED
Aperture	f/1.8	f/1.8	f/2.0	f1.8
Sensor size	1/2 inch	-	1/1.72 inch	-
Pixel size	0.8 m	-	0.8 μm	-

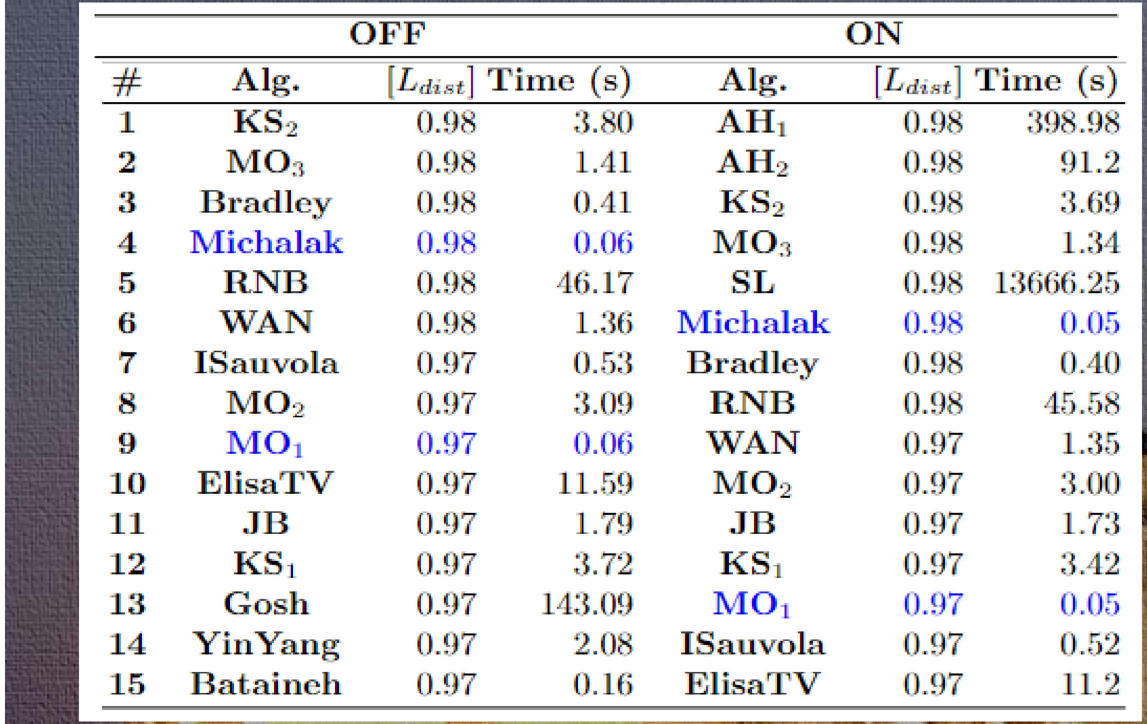
Motorola G9:

Printing



	OFF			ON		
#	Alg.	P_{err}	Time (s)	Alg.	P_{err}	Time (s)
1	Michalak	0.92	0.06	KS$_1$	0.55	3.42
2	MO$_3$	0.94	1.41	MO$_1$	0.59	0.05
3	Bradley	0.95	0.41	Gosh	0.70	145.16
4	MO$_1$	0.97	0.06	Yasin	0.74	1.75
5	ElisaTV	1.06	11.59	ElisaTV	0.83	11.2
6	Yasin	1.14	2.03	MO$_3$	0.86	1.34
7	DilatedUNet	1.17	188.27	Bradley	0.91	0.40
8	MO$_2$	1.19	3.09	Michalak	0.97	0.05
9	Gosh	1.24	143.09	Singh	1.00	0.44
10	WX	1.25	281.66	Nick	1.12	0.21
11	KS$_2$	1.42	3.80	Su-Lu	1.22	2.17
12	DocDLink	1.43	300.18	DilatedUNet	1.24	187.73
13	KS$_1$	1.68	3.72	Wolf	1.32	0.29
14	ISauvola	1.72	0.53	WX	1.64	281.16
15	Su-Lu	1.74	2.19	MO$_2$	1.65	3.00

OCR

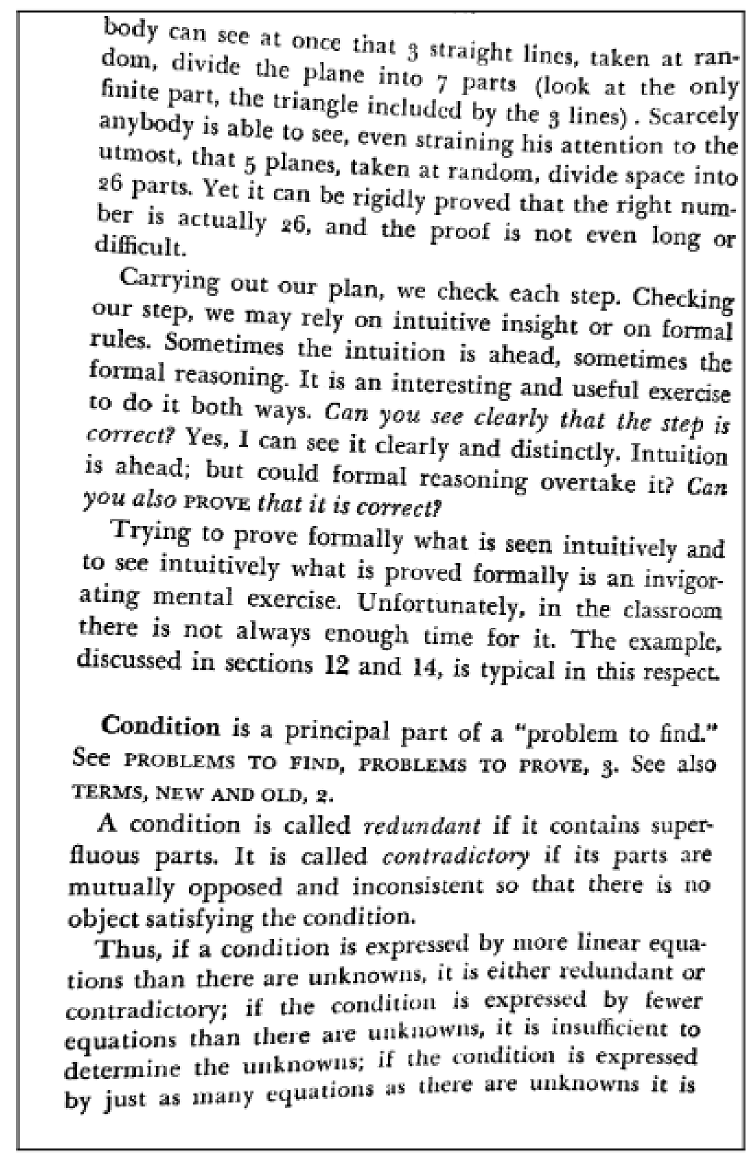


	OFF			ON		
#	Alg.	$[L_{dist}]$	Time (s)	Alg.	$[L_{dist}]$	Time (s)
1	KS$_2$	0.98	3.80	AH$_1$	0.98	398.98
2	MO$_3$	0.98	1.41	AH$_2$	0.98	91.2
3	Bradley	0.98	0.41	KS$_2$	0.98	3.69
4	Michalak	0.98	0.06	MO$_3$	0.98	1.34
5	RNB	0.98	46.17	SL	0.98	13666.25
6	WAN	0.98	1.36	Michalak	0.98	0.05
7	ISauvola	0.97	0.53	Bradley	0.98	0.40
8	MO$_2$	0.97	3.09	RNB	0.98	45.58
9	MO$_1$	0.97	0.06	WAN	0.97	1.35
10	ElisaTV	0.97	11.59	MO$_2$	0.97	3.00
11	JB	0.97	1.79	JB	0.97	1.73
12	KS$_1$	0.97	3.72	KS$_1$	0.97	3.42
13	Gosh	0.97	143.09	MO$_1$	0.97	0.05
14	YinYang	0.97	2.08	ISauvola	0.97	0.52
15	Bataineh	0.97	0.16	ElisaTV	0.97	11.2

Overall Winner: Michalak

Motorola G9: Michalak

offset printed
book page
strobeflash on



body can see at once that 3 straight lines, taken at random, divide the plane into 7 parts (look at the only finite part, the triangle included by the 3 lines). Scarcely anybody is able to see, even straining his attention to the utmost, that 5 planes, taken at random, divide space into 26 parts. Yet it can be rigidly proved that the right number is actually 26, and the proof is not even long or difficult.

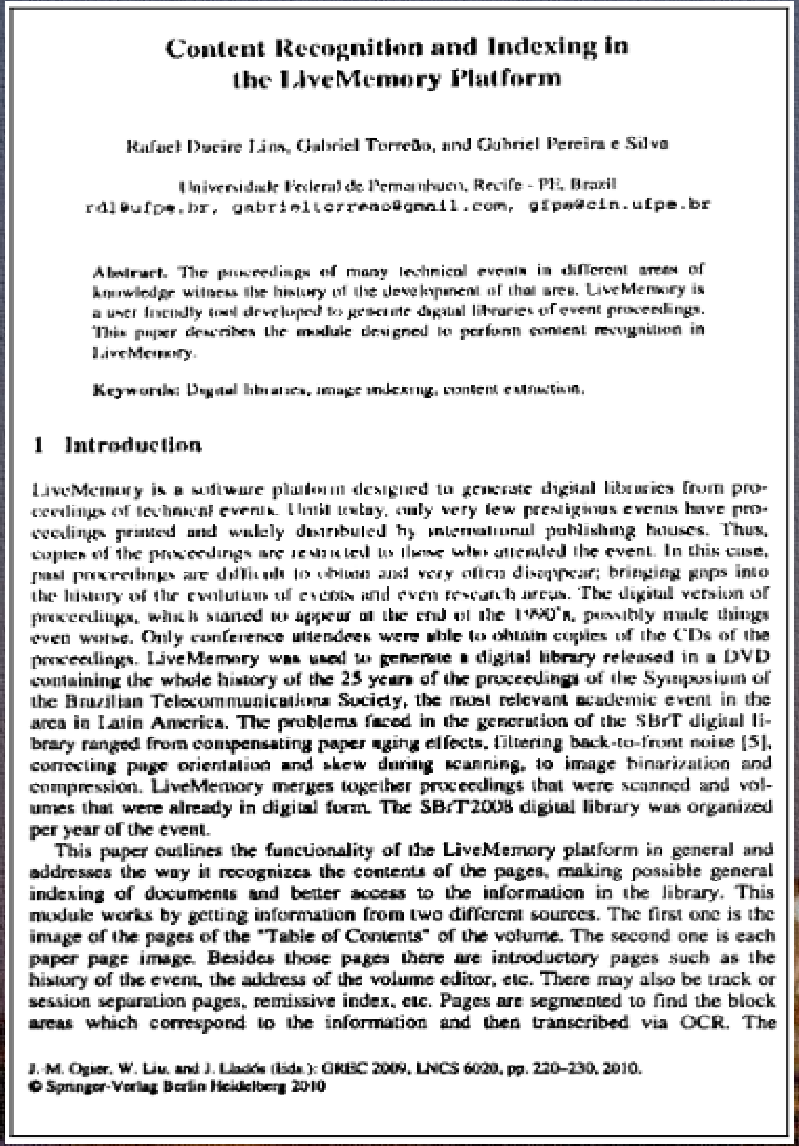
Carrying out our plan, we check each step. Checking our step, we may rely on intuitive insight or on formal rules. Sometimes the intuition is ahead, sometimes the formal reasoning. It is an interesting and useful exercise to do it both ways. *Can you see clearly that the step is correct?* Yes, I can see it clearly and distinctly. Intuition is ahead; but could formal reasoning overtake it? *Can you also* PROVE *that it is correct?*

Trying to prove formally what is seen intuitively and to see intuitively what is proved formally is an invigorating mental exercise. Unfortunately, in the classroom there is not always enough time for it. The example, discussed in sections 12 and 14, is typical in this respect.

Condition is a principal part of a "problem to find." See PROBLEMS TO FIND, PROBLEMS TO PROVE, 3. See also TERMS, NEW AND OLD, 2.

A condition is called *redundant* if it contains superfluous parts. It is called *contradictory* if its parts are mutually opposed and inconsistent so that there is no object satisfying the condition.

Thus, if a condition is expressed by more linear equations than there are unknowns, it is either redundant or contradictory; if the condition is expressed by fewer equations than there are unknowns, it is insufficient to determine the unknowns; if the condition is expressed by just as many equations as there are unknowns it is



Content Recognition and Indexing in the LiveMemory Platform

Rafael Dueire Lins, Gabriel Torreão, and Gabriel Pereira e Silva

Universidade Federal de Pernambuco, Recife - PE, Brazil
rdl@ufpe.br, gabrieltorreao@gmail.com, gfps@cin.ufpe.br

Abstract. The proceedings of many technical events in different areas of knowledge witness the history of the development of that area. LiveMemory is a user friendly tool developed to generate digital libraries of event proceedings. This paper describes the module designed to perform content recognition in LiveMemory.

Keywords: Digital libraries, image indexing, content extraction.

1 Introduction

LiveMemory is a software platform designed to generate digital libraries from proceedings of technical events. Until today, only very few prestigious events have proceedings printed and widely distributed by international publishing houses. Thus, copies of the proceedings are restricted to those who attended the event. In this case, past proceedings are difficult to obtain and very often disappear; bringing gaps into the history of the evolution of events and even research areas. The digital version of proceedings, which started to appear at the end of the 1990's, possibly made things even worse. Only conference attendees were able to obtain copies of the CDs of the proceedings. LiveMemory was used to generate a digital library released in a DVD containing the whole history of the 25 years of the proceedings of the Symposium of the Brazilian Telecommunications Society, the most relevant academic event in the area in Latin America. The problems faced in the generation of the SBrT digital library ranged from compensating paper aging effects, filtering back-to-front noise [5], correcting page orientation and skew during scanning, to image binarization and compression. LiveMemory merges together proceedings that were scanned and volumes that were already in digital form. The SBrT'2008 digital library was organized per year of the event.

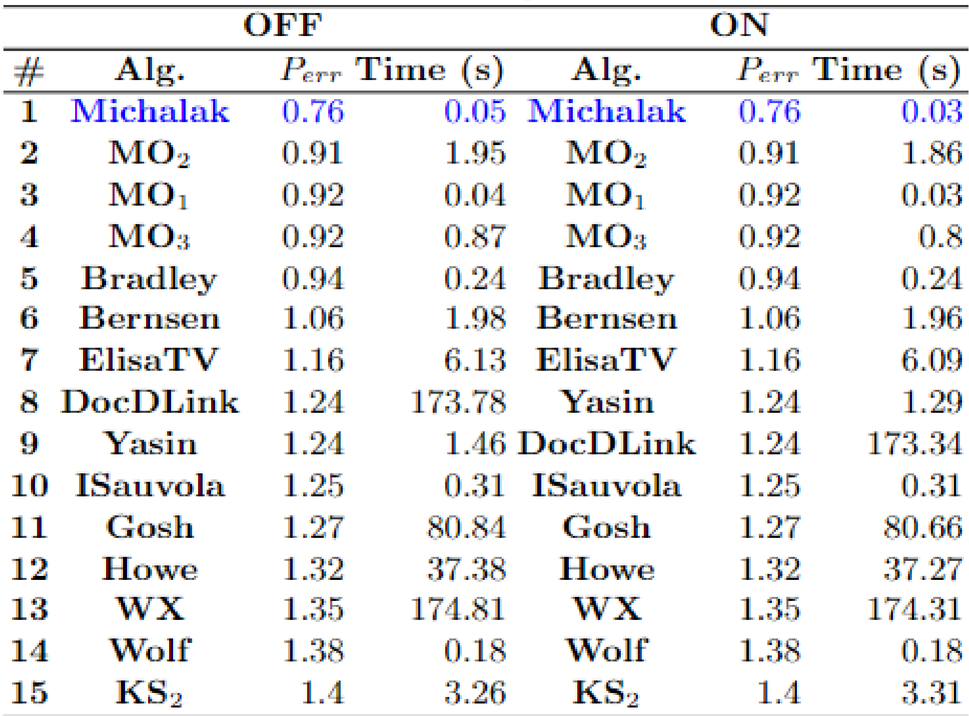
This paper outlines the funcionality of the LiveMemory platform in general and addresses the way it recognizes the contents of the pages, making possible general indexing of documents and better access to the information in the library. This module works by getting information from two different sources. The first one is the image of the pages of the "Table of Contents" of the volume. The second one is each paper page image. Besides those pages there are introductory pages such as the history of the event, the address of the volume editor, etc. There may also be track or session separation pages, remissive index, etc. Pages are segmented to find the block areas which correspond to the information and then transcribed via OCR. The

J.-M. Ogier, W. Liu, and J. Lladós (Eds.): GREC 2009, LNCS 6020, pp. 220–230, 2010.

deskjet printed
book page
strobeflash off

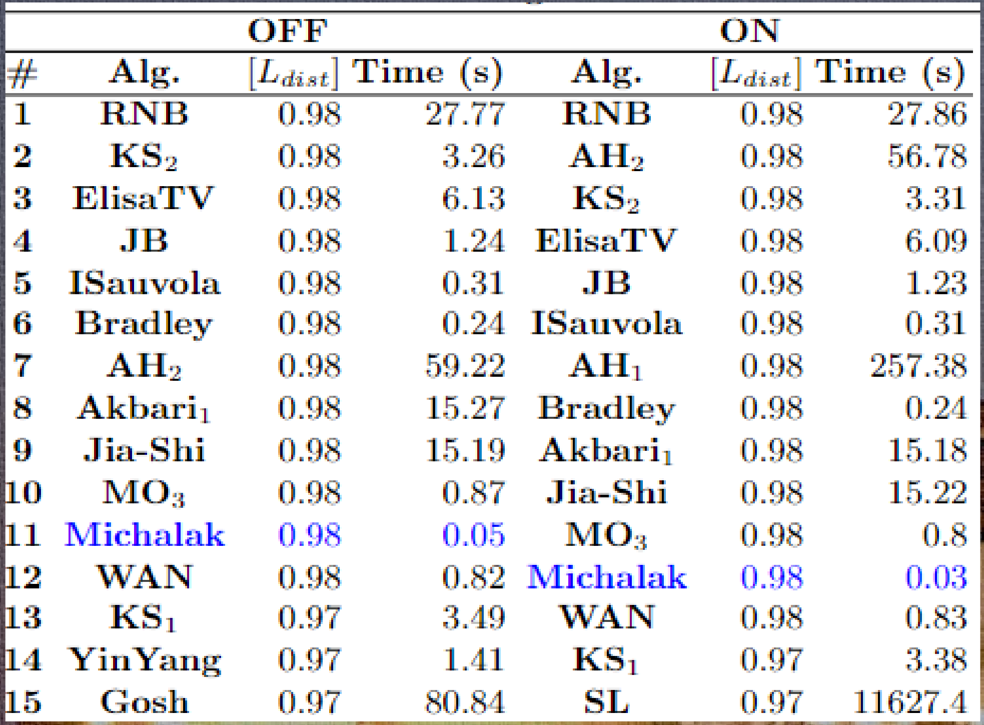
Samsung A10:

Printing



	OFF			ON		
#	Alg.	P_{err}	Time (s)	Alg.	P_{err}	Time (s)
1	Michalak	0.76	0.05	Michalak	0.76	0.03
2	MO_2	0.91	1.95	MO_2	0.91	1.86
3	MO_1	0.92	0.04	MO_1	0.92	0.03
4	MO_3	0.92	0.87	MO_3	0.92	0.8
5	Bradley	0.94	0.24	Bradley	0.94	0.24
6	Bernsen	1.06	1.98	Bernsen	1.06	1.96
7	ElisaTV	1.16	6.13	ElisaTV	1.16	6.09
8	DocDLink	1.24	173.78	Yasin	1.24	1.29
9	Yasin	1.24	1.46	DocDLink	1.24	173.34
10	ISauvola	1.25	0.31	ISauvola	1.25	0.31
11	Gosh	1.27	80.84	Gosh	1.27	80.66
12	Howe	1.32	37.38	Howe	1.32	37.27
13	WX	1.35	174.81	WX	1.35	174.31
14	Wolf	1.38	0.18	Wolf	1.38	0.18
15	KS_2	1.4	3.26	KS_2	1.4	3.31

OCR

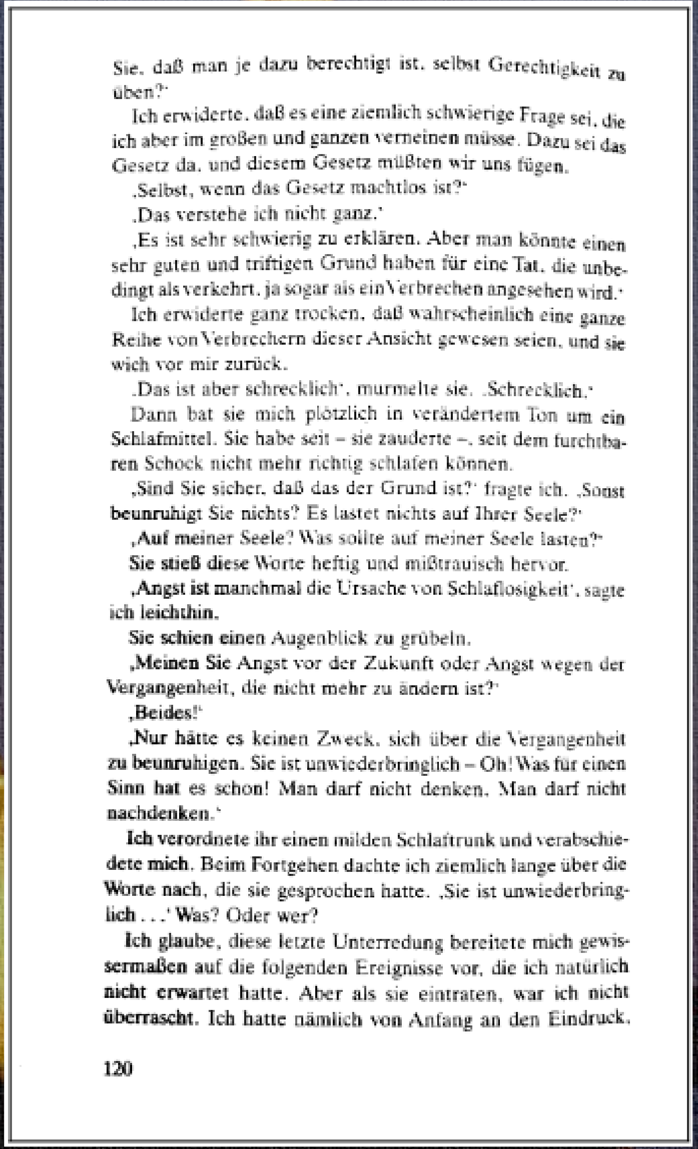


	OFF			ON		
#	Alg.	$[L_{dist}]$	Time (s)	Alg.	$[L_{dist}]$	Time (s)
1	RNB	0.98	27.77	RNB	0.98	27.86
2	KS_2	0.98	3.26	AH_2	0.98	56.78
3	ElisaTV	0.98	6.13	KS_2	0.98	3.31
4	JB	0.98	1.24	ElisaTV	0.98	6.09
5	ISauvola	0.98	0.31	JB	0.98	1.23
6	Bradley	0.98	0.24	ISauvola	0.98	0.31
7	AH_2	0.98	59.22	AH_1	0.98	257.38
8	$Akbari_1$	0.98	15.27	Bradley	0.98	0.24
9	Jia-Shi	0.98	15.19	$Akbari_1$	0.98	15.18
10	MO_3	0.98	0.87	Jia-Shi	0.98	15.22
11	Michalak	0.98	0.05	MO_3	0.98	0.8
12	WAN	0.98	0.82	Michalak	0.98	0.03
13	KS_1	0.97	3.49	WAN	0.98	0.83
14	YinYang	0.97	1.41	KS_1	0.97	3.38
15	Gosh	0.97	80.84	SL	0.97	11627.4

Overall Winner: Michalak

Samsung A10: Michalak

offset printed
book page
strobeflash on



Sie, daß man je dazu berechtigt ist, selbst Gerechtigkeit zu üben?'

Ich erwiderte, daß es eine ziemlich schwierige Frage sei, die ich aber im großen und ganzen verneinen müsse. Dazu sei das Gesetz da, und diesem Gesetz müßten wir uns fügen.

‚Selbst, wenn das Gesetz machtlos ist?'

‚Das verstehe ich nicht ganz.'

‚Es ist sehr schwierig zu erklären. Aber man könnte einen sehr guten und triftigen Grund haben für eine Tat, die unbedingt als verkehrt, ja sogar als ein Verbrechen angesehen wird.'

Ich erwiderte ganz trocken, daß wahrscheinlich eine ganze Reihe von Verbrechern dieser Ansicht gewesen seien, und sie wich vor mir zurück.

‚Das ist aber schrecklich', murmelte sie. ‚Schrecklich.'

Dann bat sie mich plötzlich in verändertem Ton um ein Schlafmittel. Sie habe seit – sie zauderte –, seit dem furchtbaren Schock nicht mehr richtig schlafen können.

‚Sind Sie sicher, daß das der Grund ist?' fragte ich. ‚Sonst beunruhigt Sie nichts? Es lastet nichts auf Ihrer Seele?'

‚Auf meiner Seele? Was sollte auf meiner Seele lasten?'

Sie stieß diese Worte heftig und mißtrauisch hervor.

‚Angst ist manchmal die Ursache von Schlaflosigkeit', sagte ich leichthin.

Sie schien einen Augenblick zu grübeln.

‚Meinen Sie Angst vor der Zukunft oder Angst wegen der Vergangenheit, die nicht mehr zu ändern ist?'

‚Beides!'

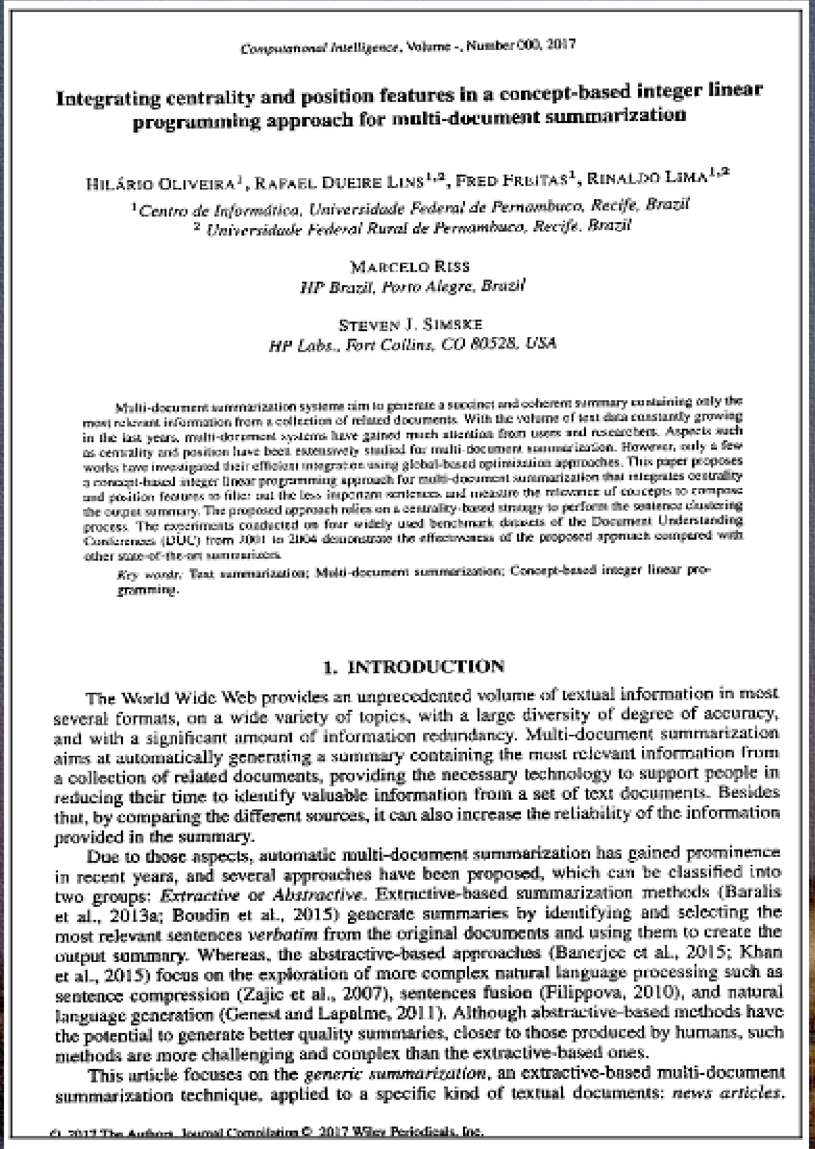
‚Nur hätte es keinen Zweck, sich über die Vergangenheit zu beunruhigen. Sie ist unwiederbringlich – Oh! Was für einen Sinn hat es schon! Man darf nicht denken. Man darf nicht nachdenken.'

Ich verordnete ihr einen milden Schlaftrunk und verabschiedete mich. Beim Fortgehen dachte ich ziemlich lange über die Worte nach, die sie gesprochen hatte. ‚Sie ist unwiederbringlich . . .' Was? Oder wer?

Ich glaube, diese letzte Unterredung bereitete mich gewissermaßen auf die folgenden Ereignisse vor, die ich natürlich nicht erwartet hatte. Aber als sie eintraten, war ich nicht überrascht. Ich hatte nämlich von Anfang an den Eindruck,

120

deskjet printed
book page
strobeflash off



Computational Intelligence, Volume -, Number 000, 2017

Integrating centrality and position features in a concept-based integer linear programming approach for multi-document summarization

HILÁRIO OLIVEIRA[1], RAFAEL DUEIRE LINS[1,2], FRED FREITAS[1], RINALDO LIMA[1,2]
[1]*Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil*
[2] *Universidade Federal Rural de Pernambuco, Recife, Brazil*

MARCELO RISS
HP Brazil, Porto Alegre, Brazil

STEVEN J. SIMSKE
HP Labs., Fort Collins, CO 80528, USA

Multi-document summarization systems aim to generate a succinct and coherent summary containing only the most relevant information from a collection of related documents. With the volume of text data constantly growing in the last years, multi-document systems have gained much attention from users and researchers. Aspects such as centrality and position have been extensively studied for multi-document summarization. However, only a few works have investigated their efficient integration using global-based optimization approaches. This paper proposes a concept-based integer linear programming approach for multi-document summarization that integrates centrality and position features to filter out the less important sentences and measure the relevance of concepts to compose the output summary. The proposed approach relies on a centrality-based strategy to perform the sentence clustering process. The experiments conducted on four widely used benchmark datasets of the Document Understanding Conferences (DUC) from 2001 to 2004 demonstrate the effectiveness of the proposed approach compared with other state-of-the-art summarizers.

Key words: Text summarization; Multi-document summarization; Concept-based integer linear programming.

1. INTRODUCTION

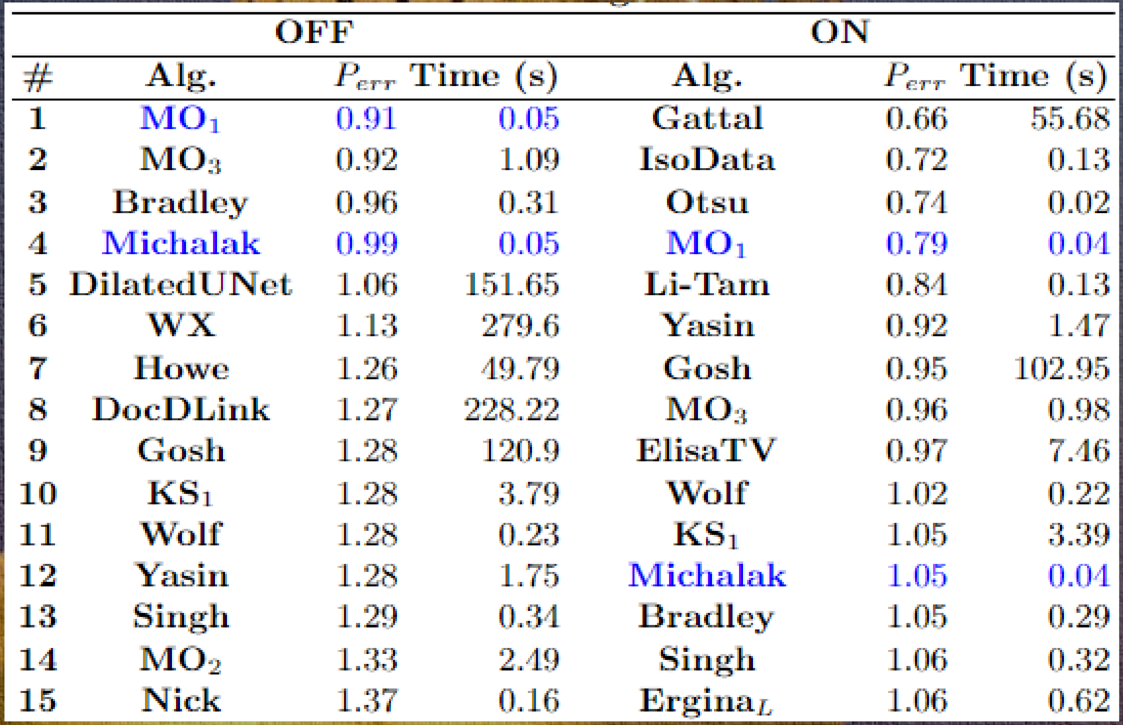
The World Wide Web provides an unprecedented volume of textual information in most several formats, on a wide variety of topics, with a large diversity of degree of accuracy, and with a significant amount of information redundancy. Multi-document summarization aims at automatically generating a summary containing the most relevant information from a collection of related documents, providing the necessary technology to support people in reducing their time to identify valuable information from a set of text documents. Besides that, by comparing the different sources, it can also increase the reliability of the information provided in the summary.

Due to those aspects, automatic multi-document summarization has gained prominence in recent years, and several approaches have been proposed, which can be classified into two groups: *Extractive or Abstractive*. Extractive-based summarization methods (Baralis et al., 2013a; Boudin et al., 2015) generate summaries by identifying and selecting the most relevant sentences *verbatim* from the original documents and using them to create the output summary. Whereas, the abstractive-based approaches (Banerjee et al., 2015; Khan et al., 2015) focus on the exploration of more complex natural language processing such as sentence compression (Zajic et al., 2007), sentences fusion (Filippova, 2010), and natural language generation (Genest and Lapalme, 2011). Although abstractive-based methods have the potential to generate better quality summaries, closer to those produced by humans, such methods are more challenging and complex than the extractive-based ones.

This article focuses on the *generic summarization*, an extractive-based multi-document summarization technique, applied to a specific kind of textual documents: *news articles*.

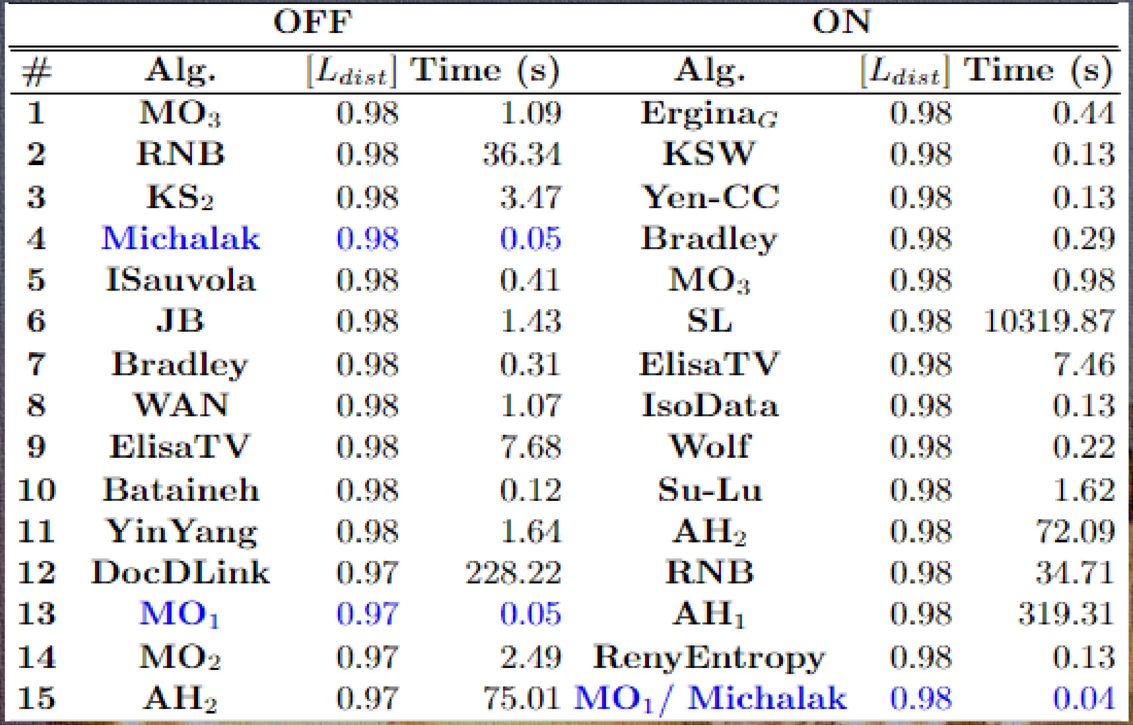
Samsung S20:

Printing



	OFF			ON		
#	Alg.	P_{err}	Time (s)	Alg.	P_{err}	Time (s)
1	MO$_1$	0.91	0.05	Gattal	0.66	55.68
2	MO$_3$	0.92	1.09	IsoData	0.72	0.13
3	Bradley	0.96	0.31	Otsu	0.74	0.02
4	Michalak	0.99	0.05	MO$_1$	0.79	0.04
5	DilatedUNet	1.06	151.65	Li-Tam	0.84	0.13
6	WX	1.13	279.6	Yasin	0.92	1.47
7	Howe	1.26	49.79	Gosh	0.95	102.95
8	DocDLink	1.27	228.22	MO$_3$	0.96	0.98
9	Gosh	1.28	120.9	ElisaTV	0.97	7.46
10	KS$_1$	1.28	3.79	Wolf	1.02	0.22
11	Wolf	1.28	0.23	KS$_1$	1.05	3.39
12	Yasin	1.28	1.75	Michalak	1.05	0.04
13	Singh	1.29	0.34	Bradley	1.05	0.29
14	MO$_2$	1.33	2.49	Singh	1.06	0.32
15	Nick	1.37	0.16	Ergina$_L$	1.06	0.62

OCR

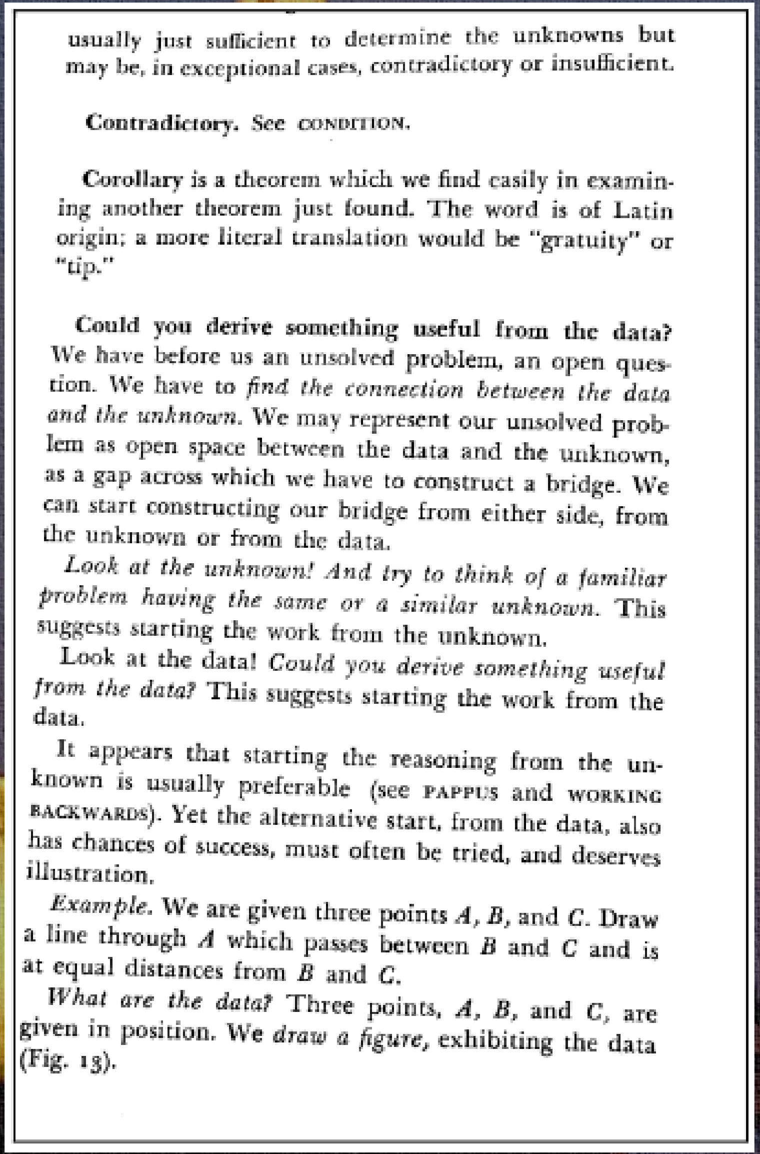


	OFF			ON		
#	Alg.	$[L_{dist}]$	Time (s)	Alg.	$[L_{dist}]$	Time (s)
1	MO$_3$	0.98	1.09	Ergina$_G$	0.98	0.44
2	RNB	0.98	36.34	KSW	0.98	0.13
3	KS$_2$	0.98	3.47	Yen-CC	0.98	0.13
4	Michalak	0.98	0.05	Bradley	0.98	0.29
5	ISauvola	0.98	0.41	MO$_3$	0.98	0.98
6	JB	0.98	1.43	SL	0.98	10319.87
7	Bradley	0.98	0.31	ElisaTV	0.98	7.46
8	WAN	0.98	1.07	IsoData	0.98	0.13
9	ElisaTV	0.98	7.68	Wolf	0.98	0.22
10	Bataineh	0.98	0.12	Su-Lu	0.98	1.62
11	YinYang	0.98	1.64	AH$_2$	0.98	72.09
12	DocDLink	0.97	228.22	RNB	0.98	34.71
13	MO$_1$	0.97	0.05	AH$_1$	0.98	319.31
14	MO$_2$	0.97	2.49	RenyEntropy	0.98	0.13
15	AH$_2$	0.97	75.01	MO$_1$/ Michalak	0.98	0.04

Overall Winner: MO$_1$ (Michalak)

Samsung S20: MO$_1$ (Michalak)

offset printed
book page
strobeflash on



usually just sufficient to determine the unknowns but may be, in exceptional cases, contradictory or insufficient.

Contradictory. See CONDITION.

Corollary is a theorem which we find easily in examining another theorem just found. The word is of Latin origin; a more literal translation would be "gratuity" or "tip."

Could you derive something useful from the data? We have before us an unsolved problem, an open question. We have to *find the connection between the data and the unknown.* We may represent our unsolved problem as open space between the data and the unknown, as a gap across which we have to construct a bridge. We can start constructing our bridge from either side, from the unknown or from the data.

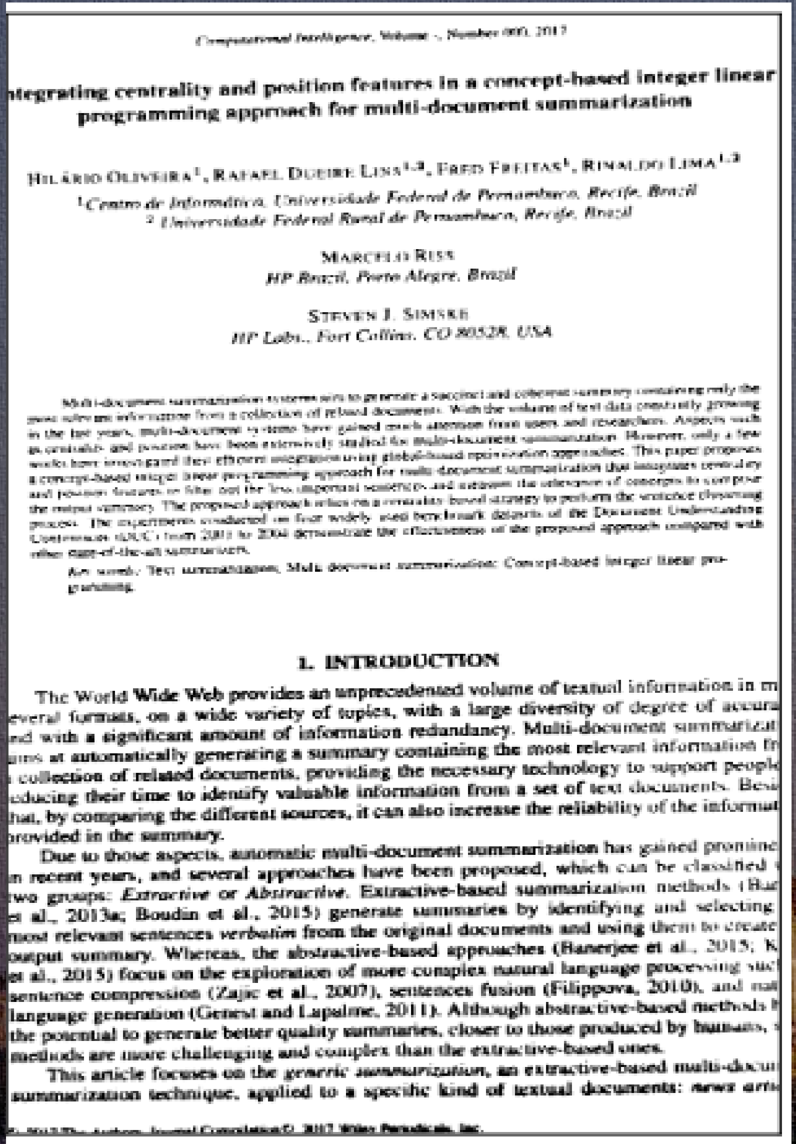
Look at the unknown! And try to think of a familiar problem having the same or a similar unknown. This suggests starting the work from the unknown.

Look at the data! *Could you derive something useful from the data?* This suggests starting the work from the data.

It appears that starting the reasoning from the unknown is usually preferable (see PAPPUS and WORKING BACKWARDS). Yet the alternative start, from the data, also has chances of success, must often be tried, and deserves illustration.

Example. We are given three points *A*, *B*, and *C*. Draw a line through *A* which passes between *B* and *C* and is at equal distances from *B* and *C*.

What are the data? Three points, *A*, *B*, and *C*, are given in position. We *draw a figure,* exhibiting the data (Fig. 13).



ntegrating centrality and position features in a concept-based integer linear programming approach for multi-document summarization

HILÁRIO OLIVEIRA[1], RAFAEL DUEIRE LINS[1,2], FRED FREITAS[1], RINALDO LIMA[1,2]

[1] *Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil*
[2] *Universidade Federal Rural de Pernambuco, Recife, Brazil*

MARCELO RISS
HP Brazil, Porto Alegre, Brazil

STEVEN J. SIMSKE
HP Labs., Fort Collins, CO 80528, USA

Multi-document summarization systems aim to generate a succinct and coherent summary containing only the most relevant information from a collection of related documents. With the volume of text data constantly growing in the last years, multi-document systems have gained much attention from users and researchers. Aspects such as centrality and position have been extensively studied for multi-document summarization. However, only a few works have investigated their efficient integration using global-based optimization approaches. This paper proposes a concept-based integer linear programming approach for multi-document summarization that integrates centrality and position features to filter out the less important sentences and measure the relevance of concepts to compose the output summary. The proposed approach relies on a centrality-based strategy to perform the sentence clustering process. The experiments conducted on four widely used benchmark datasets of the Document Understanding Conference (DUC) from 2001 to 2004 demonstrate the effectiveness of the proposed approach compared with other state-of-the-art summarizers.

Key words: Text summarization, Multi-document summarization, Concept-based integer linear programming.

1. INTRODUCTION

The World Wide Web provides an unprecedented volume of textual information in m
everal formats, on a wide variety of topics, with a large diversity of degree of accura
nd with a significant amount of information redundancy. Multi-document summarizat
ims at automatically generating a summary containing the most relevant information fr
collection of related documents, providing the necessary technology to support people
educing their time to identify valuable information from a set of text documents. Besi
hat, by comparing the different sources, it can also increase the reliability of the informat
provided in the summary.

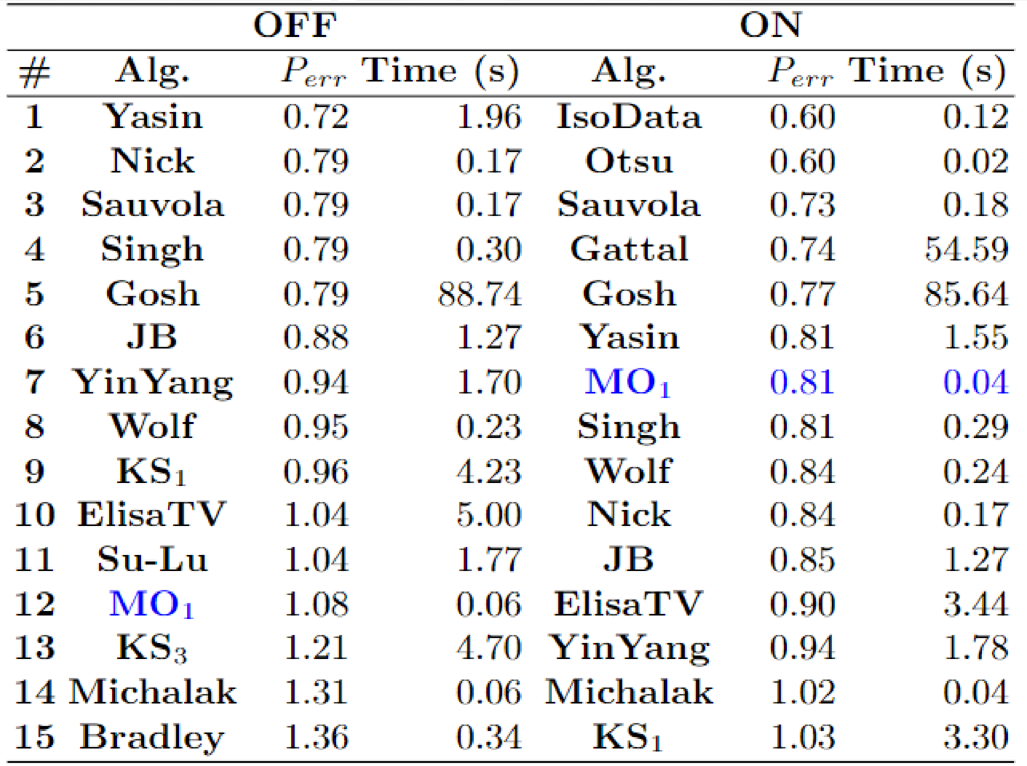
Due to those aspects, automatic multi-document summarization has gained promine
n recent years, and several approaches have been proposed, which can be classified
two groups: *Extractive* or *Abstractive*. Extractive-based summarization methods (Hu
et al., 2013a; Boudin et al., 2015) generate summaries by identifying and selecting
most relevant sentences *verbatim* from the original documents and using them to create
output summary. Whereas, the abstractive-based approaches (Banerjee et al., 2015; K
et al., 2015) focus on the exploration of more complex natural language processing suc
sentence compression (Zajic et al., 2007), sentences fusion (Filippova, 2010), and na
language generation (Genest and Lapalme, 2011). Although abstractive-based methods
the potential to generate better quality summaries, closer to those produced by humans,
methods are more challenging and complex than the extractive-based ones.

This article focuses on the *generic summarization*, an extractive-based multi-docu
summarization technique, applied to a specific kind of textual documents: *news arti*

deskjet printed
book page
strobeflash off

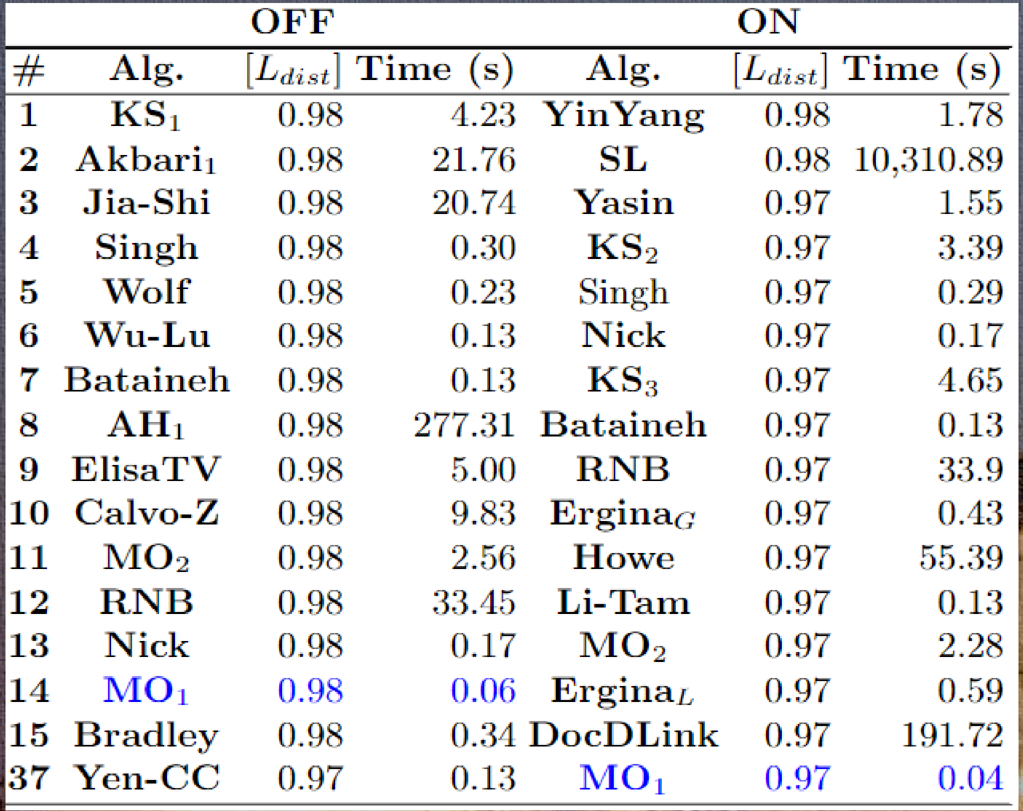
Apple Iphone SE:

Printing



	OFF			ON		
#	Alg.	P_{err}	Time (s)	Alg.	P_{err}	Time (s)
1	Yasin	0.72	1.96	IsoData	0.60	0.12
2	Nick	0.79	0.17	Otsu	0.60	0.02
3	Sauvola	0.79	0.17	Sauvola	0.73	0.18
4	Singh	0.79	0.30	Gattal	0.74	54.59
5	Gosh	0.79	88.74	Gosh	0.77	85.64
6	JB	0.88	1.27	Yasin	0.81	1.55
7	YinYang	0.94	1.70	MO$_1$	0.81	0.04
8	Wolf	0.95	0.23	Singh	0.81	0.29
9	KS$_1$	0.96	4.23	Wolf	0.84	0.24
10	ElisaTV	1.04	5.00	Nick	0.84	0.17
11	Su-Lu	1.04	1.77	JB	0.85	1.27
12	MO$_1$	1.08	0.06	ElisaTV	0.90	3.44
13	KS$_3$	1.21	4.70	YinYang	0.94	1.78
14	Michalak	1.31	0.06	Michalak	1.02	0.04
15	Bradley	1.36	0.34	KS$_1$	1.03	3.30

OCR

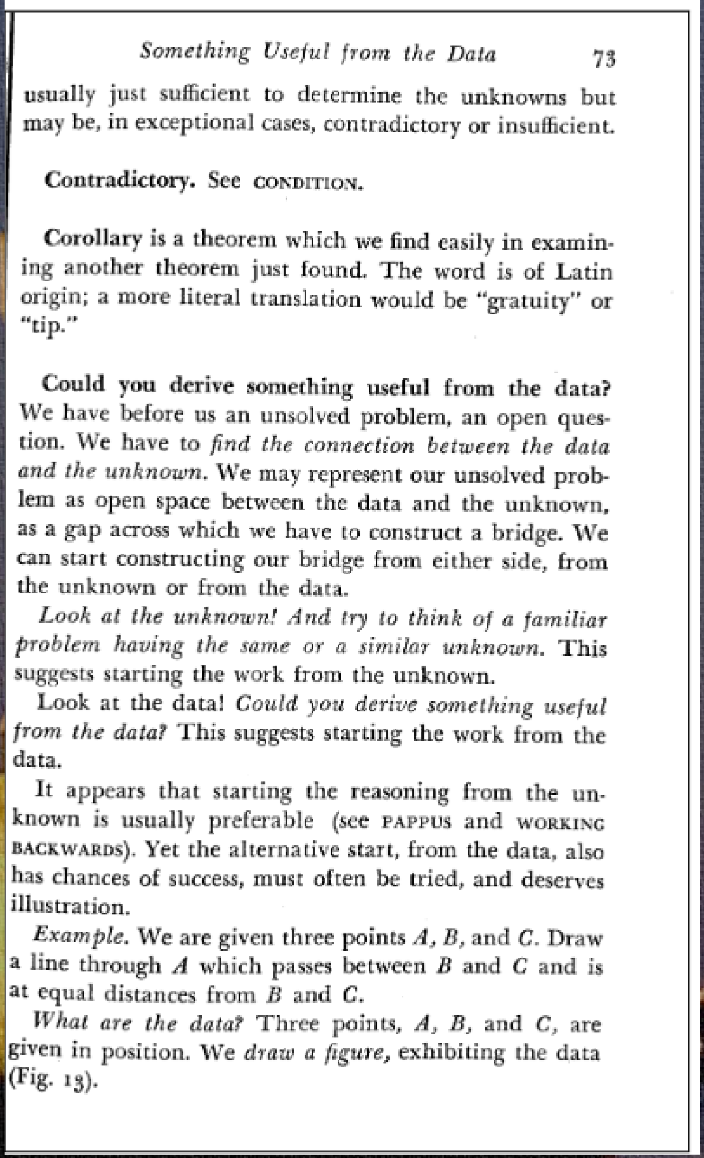


	OFF			ON		
#	Alg.	$[L_{dist}]$	Time (s)	Alg.	$[L_{dist}]$	Time (s)
1	KS$_1$	0.98	4.23	YinYang	0.98	1.78
2	Akbari$_1$	0.98	21.76	SL	0.98	10,310.89
3	Jia-Shi	0.98	20.74	Yasin	0.97	1.55
4	Singh	0.98	0.30	KS$_2$	0.97	3.39
5	Wolf	0.98	0.23	Singh	0.97	0.29
6	Wu-Lu	0.98	0.13	Nick	0.97	0.17
7	Bataineh	0.98	0.13	KS$_3$	0.97	4.65
8	AH$_1$	0.98	277.31	Bataineh	0.97	0.13
9	ElisaTV	0.98	5.00	RNB	0.97	33.9
10	Calvo-Z	0.98	9.83	Ergina$_G$	0.97	0.43
11	MO$_2$	0.98	2.56	Howe	0.97	55.39
12	RNB	0.98	33.45	Li-Tam	0.97	0.13
13	Nick	0.98	0.17	MO$_2$	0.97	2.28
14	MO$_1$	0.98	0.06	Ergina$_L$	0.97	0.59
15	Bradley	0.98	0.34	DocDLink	0.97	191.72
37	Yen-CC	0.97	0.13	MO$_1$	0.97	0.04

Overall Winner: Michalak

Apple Iphone SE: MO_1 (Michalak)

offset printed
book page
strobeflash on



usually just sufficient to determine the unknowns but may be, in exceptional cases, contradictory or insufficient.

Contradictory. See CONDITION.

Corollary is a theorem which we find easily in examining another theorem just found. The word is of Latin origin; a more literal translation would be "gratuity" or "tip."

Could you derive something useful from the data? We have before us an unsolved problem, an open question. We have to *find the connection between the data and the unknown.* We may represent our unsolved problem as open space between the data and the unknown, as a gap across which we have to construct a bridge. We can start constructing our bridge from either side, from the unknown or from the data.

Look at the unknown! And try to think of a familiar problem having the same or a similar unknown. This suggests starting the work from the unknown.

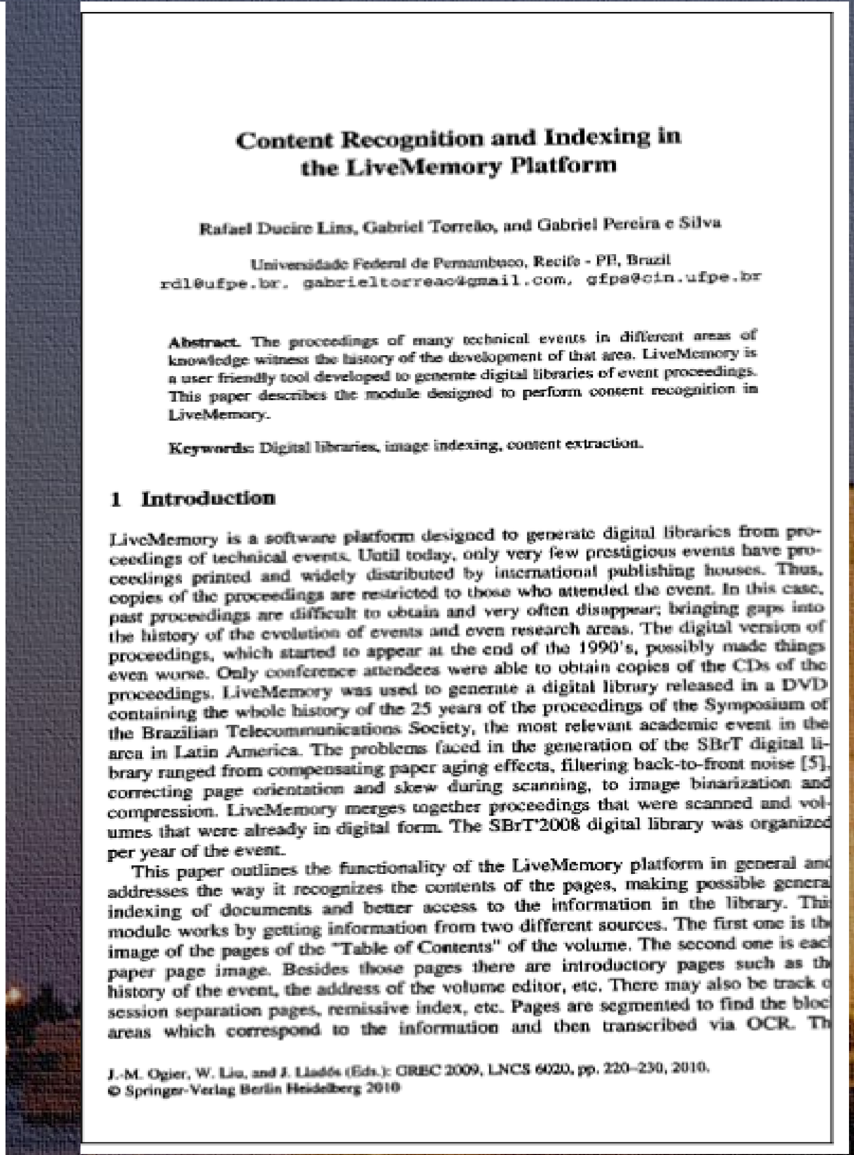
Look at the data! *Could you derive something useful from the data?* This suggests starting the work from the data.

It appears that starting the reasoning from the unknown is usually preferable (see PAPPUS and WORKING BACKWARDS). Yet the alternative start, from the data, also has chances of success, must often be tried, and deserves illustration.

Example. We are given three points *A, B,* and *C.* Draw a line through *A* which passes between *B* and *C* and is at equal distances from *B* and *C.*

What are the data? Three points, *A, B,* and *C,* are given in position. We *draw a figure,* exhibiting the data (Fig. 13).

deskjet printed
book page
strobeflash off



Content Recognition and Indexing in the LiveMemory Platform

Rafael Ducire Lins, Gabriel Torreão, and Gabriel Pereira e Silva

Universidade Federal de Pernambuco, Recife - PE, Brazil
rdl@ufpe.br, gabrieltorreao@gmail.com, gfps@cin.ufpe.br

Abstract. The proceedings of many technical events in different areas of knowledge witness the history of the development of that area. LiveMemory is a user friendly tool developed to generate digital libraries of event proceedings. This paper describes the module designed to perform content recognition in LiveMemory.

Keywords: Digital libraries, image indexing, content extraction.

1 Introduction

LiveMemory is a software platform designed to generate digital libraries from proceedings of technical events. Until today, only very few prestigious events have proceedings printed and widely distributed by international publishing houses. Thus, copies of the proceedings are restricted to those who attended the event. In this case, past proceedings are difficult to obtain and very often disappear; bringing gaps into the history of the evolution of events and even research areas. The digital version of proceedings, which started to appear at the end of the 1990's, possibly made things even worse. Only conference attendees were able to obtain copies of the CDs of the proceedings. LiveMemory was used to generate a digital library released in a DVD containing the whole history of the 25 years of the proceedings of the Symposium of the Brazilian Telecommunications Society, the most relevant academic event in the area in Latin America. The problems faced in the generation of the SBrT digital library ranged from compensating paper aging effects, filtering back-to-front noise [5], correcting page orientation and skew during scanning, to image binarization and compression. LiveMemory merges together proceedings that were scanned and volumes that were already in digital form. The SBrT'2008 digital library was organized per year of the event.

This paper outlines the functionality of the LiveMemory platform in general and addresses the way it recognizes the contents of the pages, making possible genera indexing of documents and better access to the information in the library. Thi module works by getting information from two different sources. The first one is th image of the pages of the "Table of Contents" of the volume. The second one is eac paper page image. Besides those pages there are introductory pages such as th history of the event, the address of the volume editor, etc. There may also be track o session separation pages, remissive index, etc. Pages are segmented to find the bloc areas which correspond to the information and then transcribed via OCR. Th

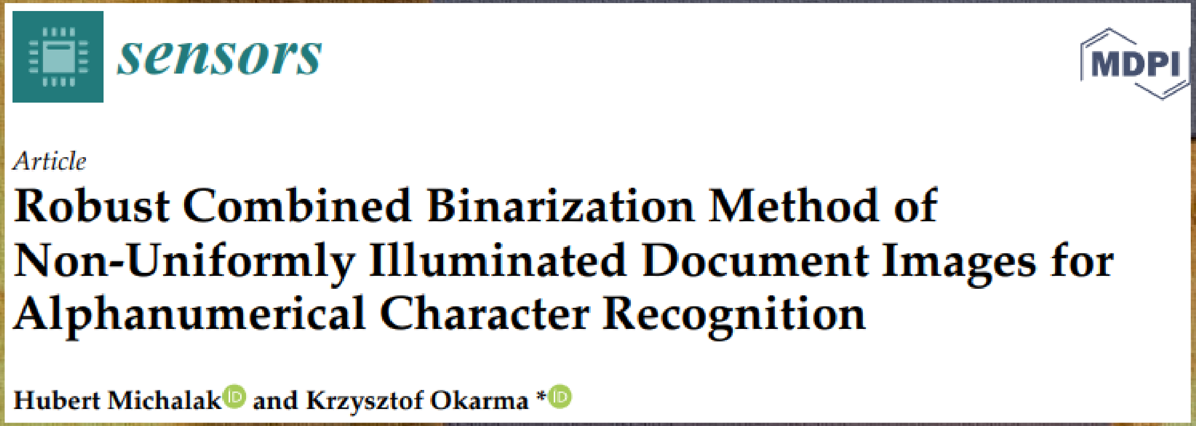
J.-M. Ogier, W. Liu, and J. Lladós (Eds.): GREC 2009, LNCS 6020, pp. 220–230, 2010.

Conclusions:

- No algorithm is good for all kinds of documents.
- The quality of the resulting image depends on the features of each image.
- Time performance is important for applicability!

Conclusions:

- Michalak and Okarma algorithms are the present 1st choice for photographed documents.



sensors

MDPI

Article

Robust Combined Binarization Method of Non-Uniformly Illuminated Document Images for Alphanumerical Character Recognition

Hubert Michalak and Krzysztof Okarma *

- This paper presents a new methodology to choose the most suitable algorithm for smartphone applications.

DocEng 2022

The ACM Symposium on Document Engineering

acm

September, 20th to 23rd San Jose, CA, USA

Quality, Space & Time Competition on Binarizing Photographed Documents

Call for competitors

Important Dates:

May 1st, 2022	Competition opens to the participants
Aug. 20th, 2022	Deadline for the registration for the contest with submission of the required executable code as well as a short description of the participants' summarization methodology.
Sep. 20th, 2022	Final contest results to be announced at the DocEng 2022 conference.

DAS 2022

15th IAPR International Workshop on Document Analysis Systems

May 22-25, La Rochelle, France

The Winner Takes It All

choosing the “best” binarization algorithm for photographed documents

Many thanks for your kind attention!

Questions?