BUILDING NON-OVERLAPPING POLYGONS FOR IMAGE DOCUMENT LAYOUT ANALYSIS RESULTS

Costin-Anton Boiangiu[1]
Mihai Zaharescu[2]
Ion Bucur[3]

ABSTRACT

Existing Computational Geometry algorithms aren't able to create a tight-fitting contour around document elements. Some of them don't follow the data points close enough, generating overlapping elements, while others, trying to keep close to the contour, break one element into pieces, falling between white spaces. The presented method both follows the contour closely and generates a single shape for a single element. The generation of the shape can be stopped at any time, either when there are no more intersections between layout elements or after a certain time elapsed, the result being valid at any given moment. In association with other algorithms it can offer a fast and clean solution for the problem of finding the non-overlapping areas resulted from the Layout Analysis document processing phase.

Keywords: layout analysis, alpha shape, bounding volumes, contour retrieval, automatic content conversion, scanned images

1. INTRODUCTION

Because of the exponential growth rate of the quantity of information, but also because of the aid computers give in document processing, document digitization has become a priority. Text is not the only information available in a document, but images, colors and layout now offer much more extra information. That is why OCR alone can't preserve the entire document, but there is also a need for preserving the look and feel. This is not straightforward neither for old documents which contain many types of nonstandard layouts and ornaments meant to beautify the page, nor for today's multi-column, multi-title and filled with pictures newspapers.

A document Layout analyzer has the role of enriching the OCR with the possibility of maintaining the aspect of the scanned document (fonts, header positions, paragraph characteristics, text and image positioning, and so on). One of the important aspects of a document layout analyzer is to separate headlines, paragraphs, images, etc. To do this

1 Associate Professor PhD, Department of Computer Science and Engineering, Faculty of Automatic Control and Computers Science, University “Politehnica” of Bucharest, Splaiul Independenţei 313, Bucharest, 060042, Romania, costin.boiangiu@cs.pub.ro
2 Engineer, Department of Computer Science and Engineering, Faculty of Automatic Control and Computers Science, University “Politehnica” of Bucharest, Splaiul Independenţei 313, Bucharest, 060042, Romania, mihai.zaharescu@cti.pub.ro
3 Associate Professor PhD, Department of Computer Science and Engineering, Faculty of Automatic Control and Computers Science, University “Politehnica” of Bucharest, Splaiul Independenţei 313, Bucharest, 060042, Romania, ion.bucur@cs.pub.ro

correctly, it has to surround each element in a way that the generated groups don't overlap [22].

2. METHODS

The methods vary from fast, simple algorithms which encapsulate elements in axis aligned bounding rectangles to complex ones which manage to follow the text or images enclosed closely [13].

The simplest algorithm finds the extreme points on the X and Y coordinates of an element in the image and uses them to generate their Bounding Rectangle or Axis Aligned Bounding Box (AABB) [1] [4].

There are a number of problems that arise here:

- the layout has to be strictly rectangular without inserted elements through text
- even small skew angles can change the boxes completely

The second problem is resolved by using Oriented Bounding Boxes (OBB), which are the minimum area rectangles, of any orientation, that include the layout element. The results don't compensate usually for the complexity [4], so they are seldom used in this domain.

The convex hull encapsulates the element in a minimum area convex polygon [11] [14] [16]. This was a broad studied problem so now we know many computational geometry algorithms that can calculate it with $O(N \cdot \log(N))$ complexity. Still, as mentioned in the introduction, the layouts are usually not so clean, containing, for example, images between columns, generating concave text paragraphs.

The Min-Max Shape algorithm [22] generates a kind of concave polygon by intersecting the results from Min-Max-X and Min-Max-Y. Min-Max-X Shape is obtained by taking the minimum and maximum X coordinates on every Y raster line. The same for Min-Max-Y but with Y axis parallel raster lines and using the min and max Y coordinates. This could be enough for most layouts, but there are exceptions. Some really odd layouts from magazines old manuscripts can contain text areas that generate a C like shape that cannot be deduced with this method.

The Alpha-Shape [19] is a generalization of the convex hull and a sub-graph of the Delaunay triangulation [12]. It can be visualized as the contour created by a sphere or radius R that is rolling onto the exterior of the points in the same direction. It can’t go over the points. The sphere will enter between points that are farther away than 2R but will roll over the ones that are closer than 2R. The result is a tight concave polygon that contains all the points that were touched by the sphere. The problem with the construction of the Alpha-Shape is that it is usually based on the Delaunay triangulation, which is not very economical. Also, there is the necessity of choosity how fine or coarse the generated polygon is, based upon how far apart are the farthest apart two neighboring points, so the sphere doesn't fall between them.

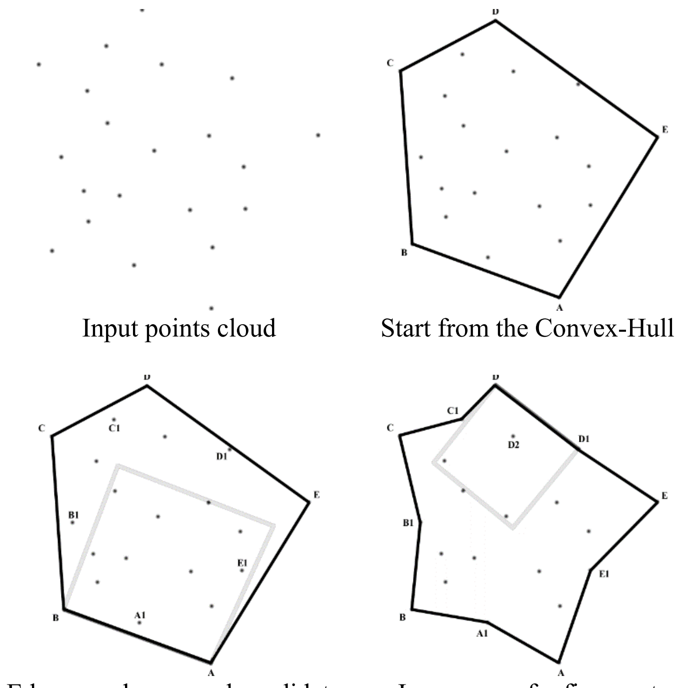
3. BETA-SHAPE

Even if Alpha-Shapes seem promising, they are unusable in the case of documents, because the distance between items, like text characters or lines inside paragraphs, is not controllable, so a radius that both generates a fine contour and doesn't split an object can possibly not exist.

That is why the algorithm described in this paper always creates only one tight concave shape [18] that contains all the input points.

The algorithm starts by generating the convex hull of a set of points using one of the known and robust computational geometry algorithms. Then it starts to iteratively refine each segment of the shape by adding new points until the distance between two adjacent points is lower than a distance given as input parameter. That segment will finally be integrated into the Beta-Shape [23] [24] and will suffer no more refinements. This distance can be considered as the input radius from the Alpha-Shape.

The points that don't make up the wrapper polygon are the candidate points. For every segment of the shape that is longer than an (*EdgeLength) threshold distance*, the algorithm searches for the closest point to the segment, from the candidates, firstly inside a search area near the segment. The search area is a square region that is inside the polygon and has one edge the segment that has to be split. This search area is used in order not to encourage the usage of potentially better candidates from other edges but also to restrict the search to a subset of the input points.



Input points cloud Start from the Convex-Hull

Edge search area and candidates In process of refinement

Figure 1. Beta-Shape stages

The initial Alpha-Shape contains all the input points inside. Because at every iteration only the closest point to each segment is inserted, there is no risk of leaving out any point or to generate self intersections. As a result, after any iteration the shape is fully valid. This means that the generation process can be stopped at any time, giving the possibility to trade between execution time and refinement quality.

As optimization, the input points can be stored inside a two-dimensional hash table, testing only the points from that hash table that contain the square search area:

Bounding Rectangle	Relation between input points: $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$			
	$x1 \le x2$		$x1 > x2$	
	$y1 \le y2$	$y1 > y2$	$y1 \le y2$	$y1 > y2$
$XMin$	$x1 - y2 + y1$	$x1$	$x2 - y2 + y1$	$x2$
$XMax$	$x2$	$x2 - y2 + y1$	$x1$	$x1 - y2 + y1$
$YMin$	$y1$	$y2$	$y1 + x2 - x1$	$y2 + x2 - x1$
$YMax$	$y2 + x2 - x1$	$y1 + x2 - x1$	$y2$	$y1$

Table 1. The coordinates of the resulted bounding rectangle used as input for the two-dimensional hash-table

The condition that a point needs to satisfy in order to be inside the search rectangle is to be delimited by the four lines that construct the square. The Beta-Shape segment:

$$ax + by + c = 0 \qquad (1)$$

passing through the 2 points $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ giving:

$$\begin{aligned} a &= y_2 - y_1 \\ b &= x_1 - x_2 \\ c &= -ax_1 + by_1 \end{aligned} \qquad (2)$$

The two perpendicular edges are:

$$\begin{aligned} a_1 &= b \\ b_1 &= -a \\ c_1 &= -bx_1 + ay_1 \end{aligned} \quad \text{and} \quad \begin{aligned} a_2 &= b \\ b_2 &= -a \\ c_2 &= -bx_2 + ay_2 \end{aligned} \qquad (3)$$

The edge parallel with the segment at distance *EdgeLength*:

$$\begin{aligned} a_3 &= a \\ b_3 &= b \\ c_3 &= c + EdgeLength \end{aligned} \tag{4}$$

Giving the constraints:

$$\begin{aligned} & b \cdot x - a \cdot y + c_1 >= 0 \\ & b \cdot x - a \cdot y + c_2 <= 0 \\ & a \cdot x + b \cdot y + c <= EdgeLength \\ & a \cdot x + b \cdot y + c >= 0 \end{aligned} \tag{5}$$

4. RESULTS

The images presented are a typical case of a justified text column, with uneven spaces between words and shorter and longer lines of text. Setting a small enough radius that enables the Alpha-Shape to be fine enough to separate the lines of text breaks the input points into clusters. The Beta-Shape with the same distance keeps the result connected and manages to cut out the white space after the shorter lines of text.

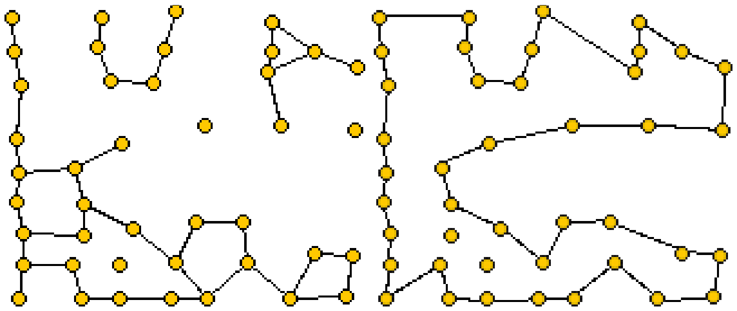


Figure 2. Alpha-Shape versus Beta-Shape

In the following test, filtered points from the player and the small image on the right are the input elements to the algorithm.

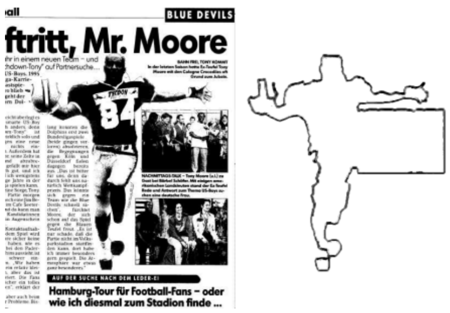

BLUE DEVILS
ftritt, Mr. Moore
Hamburg-Tour für Football-Fans – oder wie ich diesmal zum Stadion finde ...

Figure 3. Beta shape results in a “real-world” application

In this test, all the methods presented in the introduction are used, depending on the type of elements. Area 7 is the one obtained with the Beta-Shape and the algorithm kept just the points needed to stop the elements from overlapping. Segments that don't generate intersections with other layout elements don't need to be split. If more than one Beta-Shape needs to be generated for a single document, they should be done in parallel so this optimization step can occur directly at the generation stage.

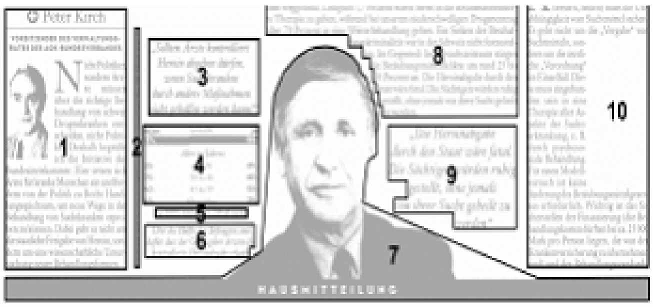


Figure 4. The Layout Analysis block-shapes as an execution result of an iterative "Non-Overlapping" algorithm

Tests show that the algorithm runs in $O(n^{\frac{5}{2}})$ for random points but in about $O(n^{\frac{3}{2}})$ for real test images [21].

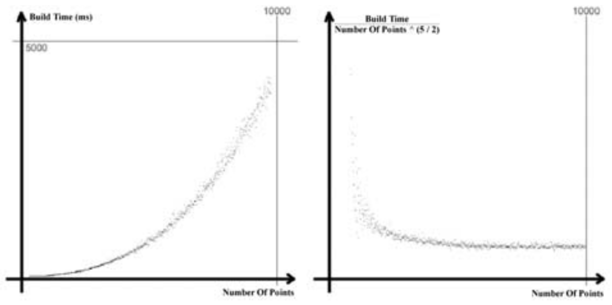


Figure 5. The complexity of Beta-Shape for randomly-distributed points

5. CONCLUSIONS

The Beta-Shape is a concept that may be successfully integrated in the Layout Analysis phase of a content conversion system because it offers a robust solution for separating document objects. It also offers valid results at every iteration step, thus being able to trade between quality and speed.

In association with other shape fitting algorithms it can be used for fast separation of the layout elements in order to accurately reconstruct the look and feel after a digitization process.

6. ACKNOWLEDGEMENTS

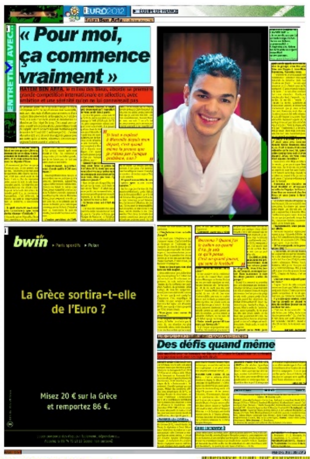
The work presented in this paper was funded by the Sectorial Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the financial agreement POSDRU/89/1.5/S/62557.

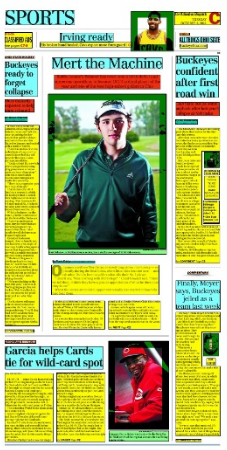
REFERENCES

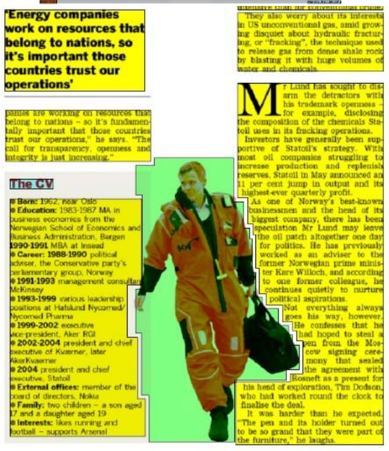
1. Douglas R. Caldwell: Unlocking the Mysteries of the Bounding Box
2. Ross S. Swick and Kenneth W. Knowles: Geographic Database Search Interfaces and the Equatorial Cylindrical Equidistant Projection
3. ESRI, 1993. Understanding GIS: The Arc/ Info method. John Wiley and Sons
4. Fast and Tight Fitting Bounding Spheres
5. J. Ritter. An efficient bounding sphere. In Andrew S. Glassner, editor, Graphics Gems. Academic Press, Boston, MA, 1990.
6. Bo Tian, Bouncing Bubble: A fast algorithm for Minimal Enclosing Ball problem 2012
7. Emo Welzl, Smallest enclosing disks (balls and ellipsoids), New Results and New Trends in Computer Science, Volume 555, 1991, pp 359-370
8. M. Bădoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. Proc. 34th Annu. ACM Sympos.on Theory of Computing, pages 250–257, 2002.
9. P. Kumar, J.S.B. Mitchell and E.A Yıldırım. Computing Core-Sets and Approximate Smallest Enclosing HyperSpheres in High Dimensions, 2003
10. K. Fischer, B. Gärtner and M. Kutz: Fast Smallest-Enclosing-Ball Computation in High Dimensions, 2003
11. Andrew, A. M. (1979), "Another efficient algorithm for convex hulls in two dimensions", Information Processing Letters 9 (5): 216–219, doi:10.1016/ 0020-0190(79)90072-3.
12. Brown, K. Q. (1979), "Voronoi diagrams from convex hulls", Information Processing Letters 9 (5): 223–228, doi:10.1016/ 0020-0190(79)90074-7.
13. de Berg, M.; van Kreveld, M.; Overmars, Mark; Schwarzkopf, O. (2000), Computational Geometry: Algorithms and Applications, Springer, pp. 2–8.
14. Chazelle, Bernard (1993), "An optimal convex hull algorithm in any fixed dimension", Discrete and Computational Geometry 10 (1): 377–409, doi:10.1007/ BF02573985.
15. Grünbaum, Branko (2003), Convex Polytopes, Graduate Texts in Mathematics (2nd ed.), Springer, ISBN 9780387004242.
16. Knuth, Donald E. (1992), Axioms and hulls, Lecture Notes in Computer Science, 606, Heidelberg: Springer-Verlag, p. ix+109, doi:10.1007/ 3-540-55611-7, ISBN 3-540-55611-7, MR 1226891.
17. Krein, M.; Šmulian, V. (1940), "On regularly convex sets in the space conjugate to a Banach space", Annals of Mathematics, 2nd ser. 41: 556–583, doi:10.2307/ 1968735, JSTOR 1968735, MR 2009.
18. Schneider, Rolf (1993), Convex bodies: The Brunn–Minkowski theory, Encyclopedia of Mathematics and its Applications, 44, Cambridge: Cambridge University Press, ISBN 0-521-35220-7, MR 1216521
19. N. Akkiraju, H. Edelsbrunner, M. Facello, P. Fu, E. P. Mucke, and C. Varela. "Alpha shapes: definition and software". In Proc. Internat. Comput. Geom. Software Workshop 1995, Minneapolis.
20. Edelsbrunner, Herbert (1995), "Smooth surfaces for multi-scale shape representation", Foundations of software technology and theoretical computer science (Bangalore, 1995), Lecture Notes in Comput. Sci., 1026, Berlin: Springer, pp. 391–412, MR 1458090.

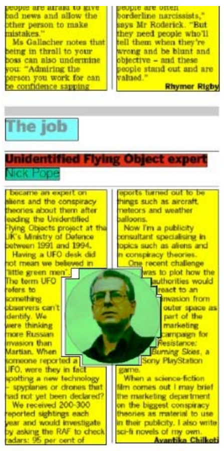
21. Edelsbrunner, Herbert; Kirkpatrick, David G.; Seidel, Raimund (1983), "On the shape of a set of points in the plane", IEEE Transactions on Information Theory 29 (4): 551–559, doi:10.1109/ TIT.1983.1056714
22. Costin-Anton Boiangiu, Bogdan Raducanu, Serban Petrescu, Ion Bucur, "Ensure Non-Overlapping in Document Layout Analysis", 20th DAAAM World Symposium, ISBN: 978-3-901509-70-4; ISSN: 1726-9679, pp.327-328, Austria Center Vienna (ACV), 25-28th of November 2009
23. Costin-Anton Boiangiu, Ion Bucur, Andrei Iulian Dvornic. "The Beta-Star Shape Algorithm for Document Clipping". Annals of DAAAM for 2008 & Proceedings of the 19th International DAAAM Symposium, Editor B. Katalinic, Published by DAAAM International (Vienna, Austria), pp. 0127–0128, Trnava, Slovakia, October 22-25, 2008, ISBN 978-3-901509-68-1, ISSN 1726-9679
24. Costin-Anton Boiangiu, "The Beta-Shape Algorithm for Polygonal Contour Reconstruction", CSCS14 – "The 14th International Conference on Control System and Computer Science", Bucharest, Romania, July 2003.

ANNEX 1 – COMPLEX DOCUMENT LAYOUTS AND THE RESULTED "NON-OVERLAPPED" SHAPES


« Pour moi, ça commence vraiment »
bwin
La Grèce sortira-t-elle de l'Euro ?
Misez 20 € sur la Grèce et remportez 86 €.
Des défis quand même


SPORTS
Irving ready
Mert the Machine
Buckeyes ready to forget collapse
Buckeyes confident after first road win
Finally, Meyer says, Buckeyes jelled as a team last week
Garcia helps Cards tie for wild-card spot


'Energy companies work on resources that belong to nations, so it's important those countries trust our operations'
The CV



The Christian Science Monitor

EDITOR'S VIEW

The modern Olympics, for better or worse, are set up as competitions among nations. So sometimes the men's Olympic basketball champions, for example, are a team that could probably not have beaten the winner of the NBA Finals. What if the Miami Heat were simply sent on to the Games, whether all the players were Americans or not? Would that be better or worse??

Stoking nationalism has long been a criticism of the Games. Europeans are especially sensitive to the negative side of nationalism – from German supremacism in the 1930s to Serbian ethnic cleansing in the 1990s. Preventing wars stoked by nationalism is a large part of what the European Union is all about.

But the Olympics-by-nation has a leveling effect as well. The massive Chinese and American teams march into the ceremonies the same way as the team from tiny Tuvalu. Hearts stir to Ireland's first-ever Olympic gymnast, who was once told nerve damage might keep him from ever walking again.

The Olympics are not just a rivalry but a celebration of breathtaking, superlative excellence, whatever country it comes from. In that spirit, the Games celebrate national differences and commonalities too.

Marshall Ingwerson,
Managing editor

DAILY NEWS BRIEFING

July 30, 2012

Romney: in Israel, on foreign policy

CANDIDATE CALLS IRAN HIS 'HIGHEST NATIONAL SECURITY PRIORITY.'

By JOSHUA MITNICK
CORRESPONDENT

[illegible] – In a signal foreign policy speech in Jerusalem, former Massachusetts Gov. Mitt Romney said that preventing Iran from building a nuclear weapon should be the "highest national security priority," and chided President Obama for quarreling with Prime Minister Benjamin Netanyahu over how to do it.

However, the presumptive Republican nominee for president offered few hints on what he would do differently from the Obama administration. Though a senior adviser, Dan Senor, earlier suggested that Mr. Romney would "respect" an Israeli decision to launch a lone attack, Romney in his speech reiterated the more vague formulation of the Obama administration that "we recognize Israel's right to defend itself."

He did reproach the president indirectly for chastising the Israeli leaders for saber rattling, a remark made when Mr. Netanyahu last visited Washington. "It is sometimes said that those who are the most committed to stopping the Iranian regime from securing nuclear weapons are reckless and provocative and inviting war," Romney said. "The opposite is true. We are the true peacemakers."

Romney's visit has escalated the competition between the two campaigns to demonstrate who's a stronger supporter of the Jewish state. Mr. Obama on Friday announced $70 million in additional military aid to help develop a rocket system to intercept missiles from Gaza and Lebanon.

Israeli officials and analysts are uneasy with Israel being in the election's spotlight. That's because an election debate over Israel could damage long-term ties between the countries by risking the Jewish state's long-cultivated bipartisan support in Washington.

▷ To read the full article, CLICK HERE

This is today's Monitor top story. For more of today's stories, go to CSMonitor.com.

Developing countries go mobile

CELLPHONE USE RESHAPES BILLIONS OF LIVES, WORLD BANK REPORTS.

By WHITNEY EULICH
STAFF WRITER

From remote farms to rural health centers, one thing is transforming how even the world's poorest people live: the mobile phone.

Cellphone use in the developing world has climbed to nearly 5 billion mobile subscriptions, and three-quarters of the world now has access to mobile networks.

This technology is reshaping the way individuals and communities manage their finances, monitor weather, engage with government, and earn a living, according to the recent World Bank Maximizing Mobile report.

"People are going from zero to 60. It is huge to go from no phone at all to a cellphone," says Anne Nelson, international media development specialist and adjunct professor at Columbia University. "The rapid penetration of cellphones in developing countries is changing lives dramatically."

Mobile devices in regions like Africa are largely limited to voice and Short Message Service texting, but even the most basic mobile communications can increase school attendance, facilitate banking or cash transfers, create jobs, measure health indicators, accelerate disaster response, and fuel citizen engagement in governance and democracy.

For example, in Niger, access to cellphones has allowed grain traders to compare market prices across the country, cutting the cost of traveling to different markets and resulting in profit improvements of nearly 30 percent for traders.

According to the World Bank, as phones become cheaper and more robust and networks double their bandwidth almost every year and a half, more people have access to mobile devices than they do to clean drinking water or electricity in some countries.

▷ To read the full article, CLICK HERE

© 2012 The Christian Science Publishing Society
ISSN [illegible]

JULY 30, 2012 THE CHRISTIAN SCIENCE MONITOR

MENSCHEN & WIRTSCHAFT

Fahrradhändler vertreibt auch Möbel

René Marius Köhler hat 14 Onlineshops gegründet und beschäftigt 150 Mitarbeiter

[illegible]

René Marius Köhler

Die Gründer

[illegible]

OLIVER SCHMALE

[illegible]

Fräulein Dr. Vera Möbius und René Dubot, Öl-Ingenieur, kämpfen um ein Stückchen Wüste. Dort liegt sie, die geheimnisvolle, versunkene „Stadt der Hellen Brunnen". Es geht um Forscherruhm und Öl, und es wird scharf geschossen.

Küsse gegen Dynamit

ROMAN VON JOHANNES ANDRÉ

[illegible]

NACHRICHTEN

Eine Stadt bangt um einen Arzt, der keiner ist

[illegible]

Rennfahrer Richard von Frankenberg:

Keine Angst vor großem Tempo!

[illegible]

Wo ist unser Kind!

[illegible]