Towards Standardized Vectorial Resource Descriptors on the Web

Wolfgang Orthuber
Department of Orthodontics
University Clinic Schleswig-Holstein
Arnold Heller Str. 16
24105 Kiel / Germany
orthuber@kfo-zmk.uni-kiel.de

Stefan Dietze
Knowledge Media Institute
The Open University
MK7 6AA
Milton Keynes / United Kingdom
s.dietze@open.ac.uk

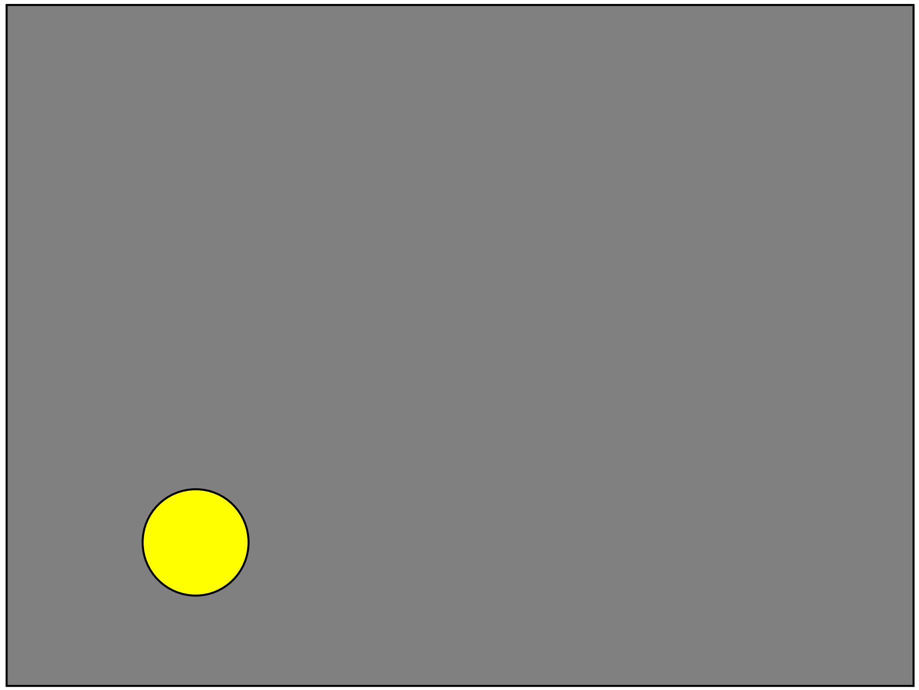
Example

Aim: search resources with certain quantitative properties,

e.g. search a resource with a position

Describe position

using language!



Example

Word based description

Position =

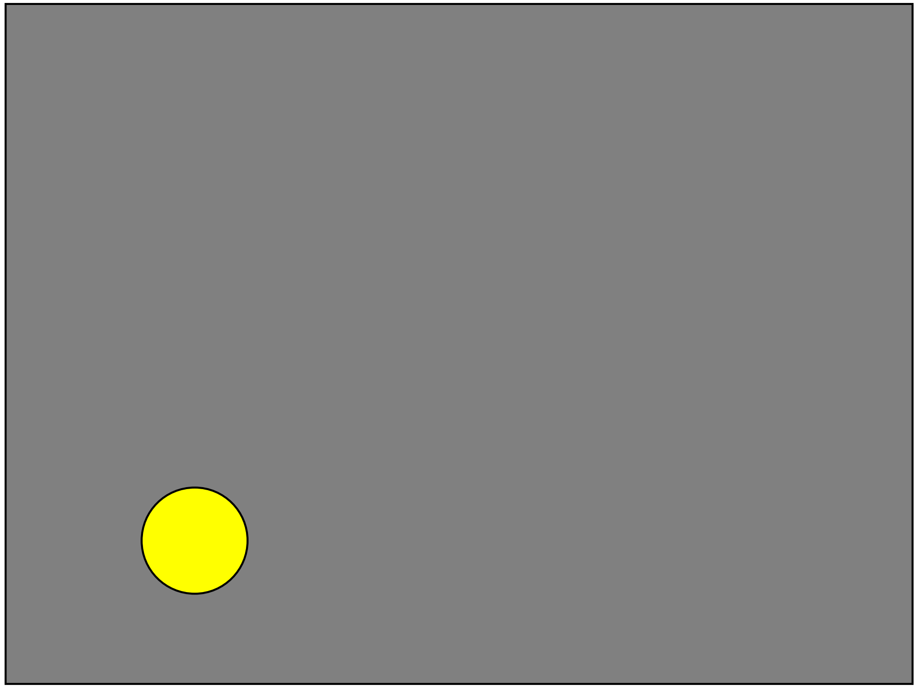
„near the left margin,

but not very near,

and close to the bottom,

but not very close ...“

Not searchable!



Example

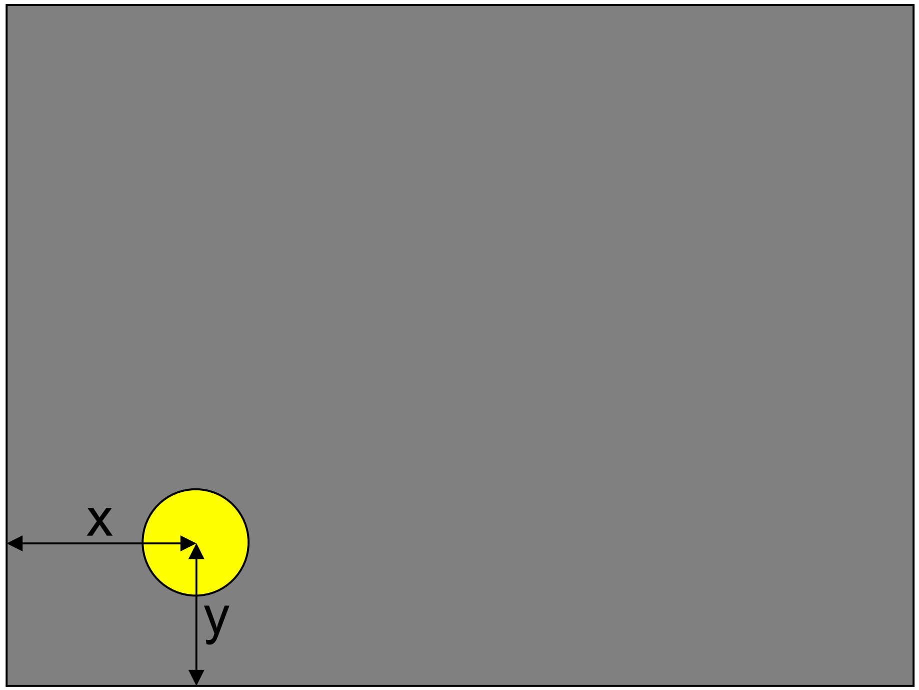
Vectorial representation

Position = (x,y)

(x,y) is a vector;

vectors are numerically searchable:

- isolate comparable vectors on the web
- calculate distances to the searched vector
- sort result: the smaller the distance, the higher rank



Prerequisite for vectorial representation

- Prerequisite for vectorial representation is quantification.
- Instances of a quantitative property can be ordered from "little" to "great“.
- So a numeric representation, which inherits the original order, is possible.
- So neighboring instances can be mapped to neighboring numbers. (i.e. similar” instances can be mapped to “similar” numbers)

Even the mapping non existent → 0, existent → 1 means quantification, so “everything possible “ can have a vectorial representation. In many applications, however, fine distinction is desirable.

Aim for vectorial representation

1. Find most important (i.e. decision relevant) ordered properties of a resource, so that most important variants of the resource can be represented by as few as possible numbers and

2. Small changes of the resource are mapped to small changes of the numbers (no zigzag).

3. Group all variants of the resource by giving them a name (the VSI or Vector Space Identifier).

Symbolic versus Vectorial approach

- Human beings can keep symbols (names) better in mind than vectors.
- There are, however, much more vectors than symbols.
- Symbolic description cannot achieve the resolution and precision of vectorial description.

therefore we suggest to use

- Symbols resp. names (VSI) for rough grouping and Vectors for fine description (within the group)

A precise and comparable descriptor of everything describable:

Vectorial Resource Descriptor (VRD)

Identifier of the Quantifiable Resource (QRI)

(HTTP URI)

Vector Space Identifier (VSI)

(HTTP URI)

Feature Vector

(a sequence of numbers)

Auxiliary data

(e.g. date, keywords)

Task sharing

VSI = http URI

The VSI uniquely identifies the kind of a **VRD.** It is a http URI which (like a URL) permanently refers to a unique, permanent web address and which differs if and only if the web address differs. So it is a unique name and also a unique reference.

It points to the **Vector Space Descriptor (VSD)** which contains all defining and further associated information. **This guarantees also that there is exactly one determining definition (anchor) for this name, and indirectly well defined task sharing among all domain name owners for definition of VRDs.**

VRD

- QRI (HTTP URI)
- VSI (HTTP URI)
- Feature vector
- Auxiliary data

VSD

- Defining information
- Keywords
- Templates for human readable representation
- Metric for similarity search
- Further associated information

VRD syntax example

VRDs can be written in RDF.

Example:

```
<?xml version="1.0"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:pa="http://a.com/syntax/VRD#">
<rdf:Description rdf:about="(1)">
<pa:VSI>(2)</pa:VSI>
<dc:date>(3)</dc:date>
<dc:subject>(4)</dc:subject>
<pa:OrgData>(5)</pa:OrgData>
<pa:FVec>(6)</pa:FVec>
</rdf:Description>
</rdf:RDF>
```

Abbreviations:

(1) QRI: HTTP URI of resource whose quantitative properties are represented by feature vector (6)

(2) VSI: HTTP URI, vector space identifier and pointer to the VSD which describes the vector space (with definition of dimensions, metric etc.)

(3) date of creation of this VRD

(4) keywords, separated by commas

(5) optional URL of a file which contains original data (e.g. a picture, sound, melody, movie, sensor data) from which the feature vector has been calculated

(6) feature vector, usually a sequence of numbers which represents quantitative properties of resource (1)

Vector Space Descriptors (VSDs)

The VSI points to the VSD:

```
<?xml version="1.0"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:pa="http://a.com/syntax/VSD#">
 <rdf:Description rdf:about="(1)">
  <pa:Dimensionality>(2)</pa:Dimensionality>
  <pa:Metric>(3)</pa:Metric>
  <pa:FWeight>(4)</pa:FWeight>
  <pa:WMode>(5)</pa:WMode>
  <pa:FVecDef>(6)</pa:FVecDef>
  <pa:ViewTemplate>(7)</pa:ViewTemplate>
  <dc:subject>(8)</dc:subject>
  <dc:date>(9)</dc:date>
 </rdf:Description>
 </rdf:RDF>
```

Abbreviations:

(1) VSI: name of vector space which is described here, as HTTP URI also pointer to this file
(2) dimensionality of the vector space
(3) number of metric (distance function)
- 1: (1 - correlation coefficient)
- 2: GPS metric
- 3: discrete metric
- 4: weighted Manhattan metric
- 5: weighted Euclidean distance

(4) optional weighting vector default for every dimension: 1
(5) optional weighting mode
- 0: weighting vector is used as it is (default)
- 1: every coordinate i of the weighting vector is multiplied by 1 / (standard deviation of dimension i) before usage

(6) Complete definition and description of the vector space
(7) optional template for human readable representation
(8) optional keywords, separated by commas
(9) date of definition of this vector space

Vectorial Resource Descriptors (VRDs)

Advantages:

- Can become a well defined precise fundament of descriptions.
- Simple structured machine readable data
  (VSI + feature vector)
- As quickly as possible identifiable by the VSI.
- The feature vector can describe reality objectively and more precise and simultaneously more broadly than language.
- The search is more specific, numerical search can close gaps in language based approaches (e.g. precisely describe sounds, pictures, geometry, multidimensional diagnostic data etc.)
- Complex hierarchical structures could be mapped to VSIs.
- VRDs can be linked together with original data whose features they describe.
- →

Vectorial Resource Descriptors (VRDs)

Advantages (continuation):

- VRDs are globally interoperable, due to the unique VSI.
- VRDs are human readable using templates in the VSDs.
- VRDs are decentrally defineable by all interested domain name owners, which implies much more working capacity for definion and creation than central definition (e.g. by manufacturer of diagnostic devices).
- The VRDs can be used for advanced evaluation and modelling due to their machine readability. Details would exceed the scope of this presentation.
- →

Vectorial Resource Descriptors (VRDs)

- Due to their precision VRDs can be used to find decision relevant information, i.e. they can be used for decision support, because

VRDs are numerically searchable.

- During feature extraction similarities of original data are mapped to spatial similarities of the feature vectors. As metric a compact distance function is appropriate.

- The Manhattan distance is an example:

Let a, b denote two feature vectors with: $a = (a_1, a_2, \ldots, a_n)$, $b = (b_1, b_2, \ldots, b_n)$

Then the Manhattan distance is: $d(a-b) = w_1|a_1-b_1| + w_2|a_2-b_2| + \ldots + w_n|a_n-b_n|$

It is possible to combine distance functions.

Default weighting factor: $w_i = 1 /$ **standard deviation** *(dimension$_i$).*

Vectorial Web Search

Vectorial Web Search consists of the following steps:

- User provides a VRD or only its feature vector V with VSI.
- User confines search by a regular expression and/or by a conventional word based search string S (optional)
- Search engine selects all VRDs with the chosen VSI, optionally with string S at the associated resource (identified by QRI)
- If a regular expression is given, the collection is confined so that it fulfills this expression.
- Using the metric provided in the VSD (figure 1) the search engine calculates distances between the feature vector V and the feature vectors of the collected VRDs and sorts them according to distance
- In the search result the rank of collected VRDs and associated resources is the higher, the smaller the distance is.

Vectorial Web Search

Important limitations:

- Practically:
  - Not enough data in database (relevant especially in the beginning)
  - Inefficient or redundant definitions of vector spaces. Therefore the VSD contains keywords, so that is possible to search for existing definitions to a topic before making a new definition.

- Technically:
  - Performance. The costs of hardware for parallel processing can become relevant in case of many users and similarity search over large collections of data.
  - Curse of dimensionality. Low dimensional vector spaces are preferable because they can be handled much more efficiently than high dimensional spaces. It can become necessary that the search engine introduces an upper limit of dimensionality.

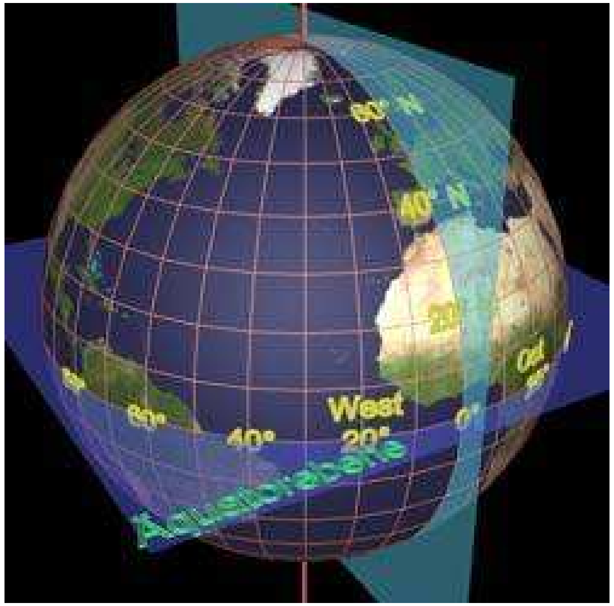
VRD Examples

GPS-Coordinates

Feature Vector:

a_1 = Latitude

a_2 = Longitude



This definition would allow combination of conventional text search with radial search.

VRD Examples

Price/d

Feature Vector:

a_1 = Price in Dollar

Price/e

Feature Vector:

a_1 = Price in Euro

Can be used to search for objects which are for sale.

VRD Examples

lwh: length, width, height

An example of a search string may be

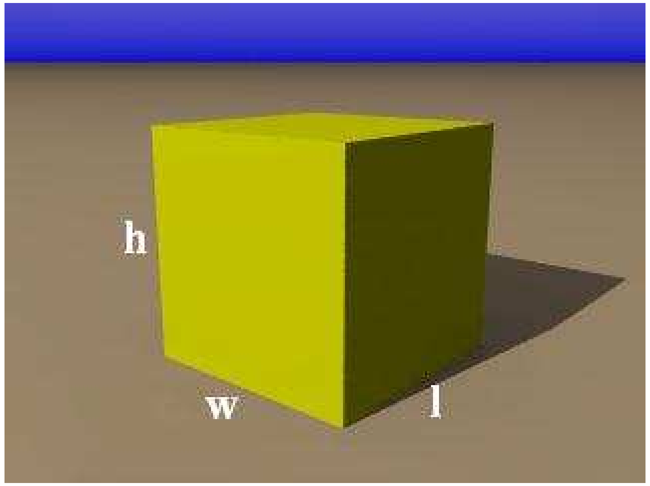
box #dn/lwh.htm 3 2 1#

The search string

box #dn/lwh.htm 3 ? ? #

or synonymously

box #dn/lwh.htm 3#



can be used to search for boxes with 3 m length and variable width and height

VRD Examples

Industrial products, e.g. electric motors 230V AC

dn/emo/230ac.htm

Feature Vector:

a_1 = power (in Watt)

a_2 = rpm (revolutions per minute)

a_3 = energy efficiency (in percent)

a_4 = axial dimeter in mm

a_5 = length in mm

a_6 = height in mm

a_7 = weight in kg

The search string # dn/emo/230ac.htm 1000 3000 #
can be used to search for 230 V AC electric Motors with 1000 W Power and 3000 rpm (revolutions per minute).

VRD Examples

Customized clothes

dn/body/size.upper

Feature Vector:

a_1 = collar size (in cm)

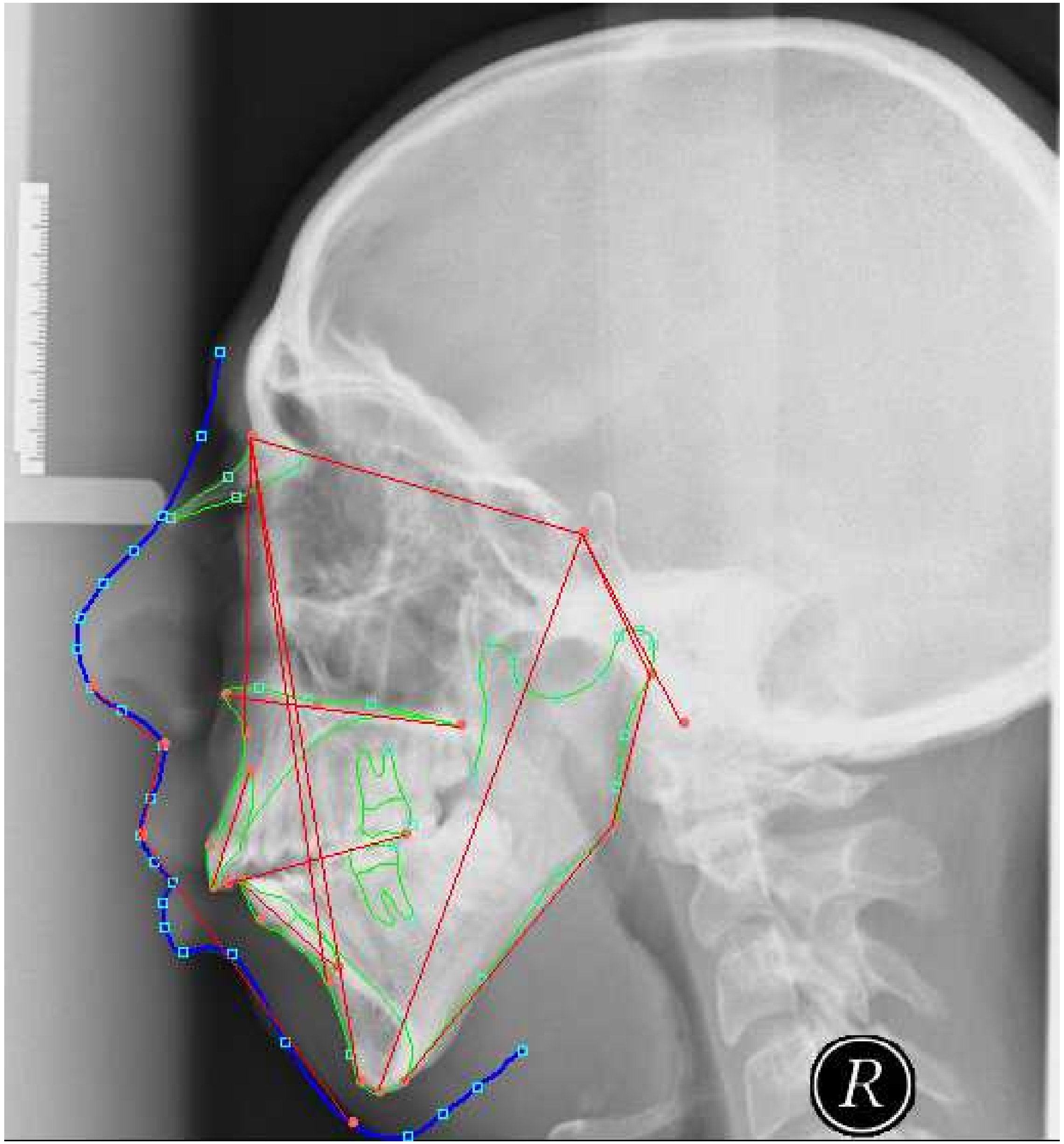
a_2 = abdominal girth (in cm)

a_3 = chest measurement (in cm)

...

This VRD be also used for ordering clothes.

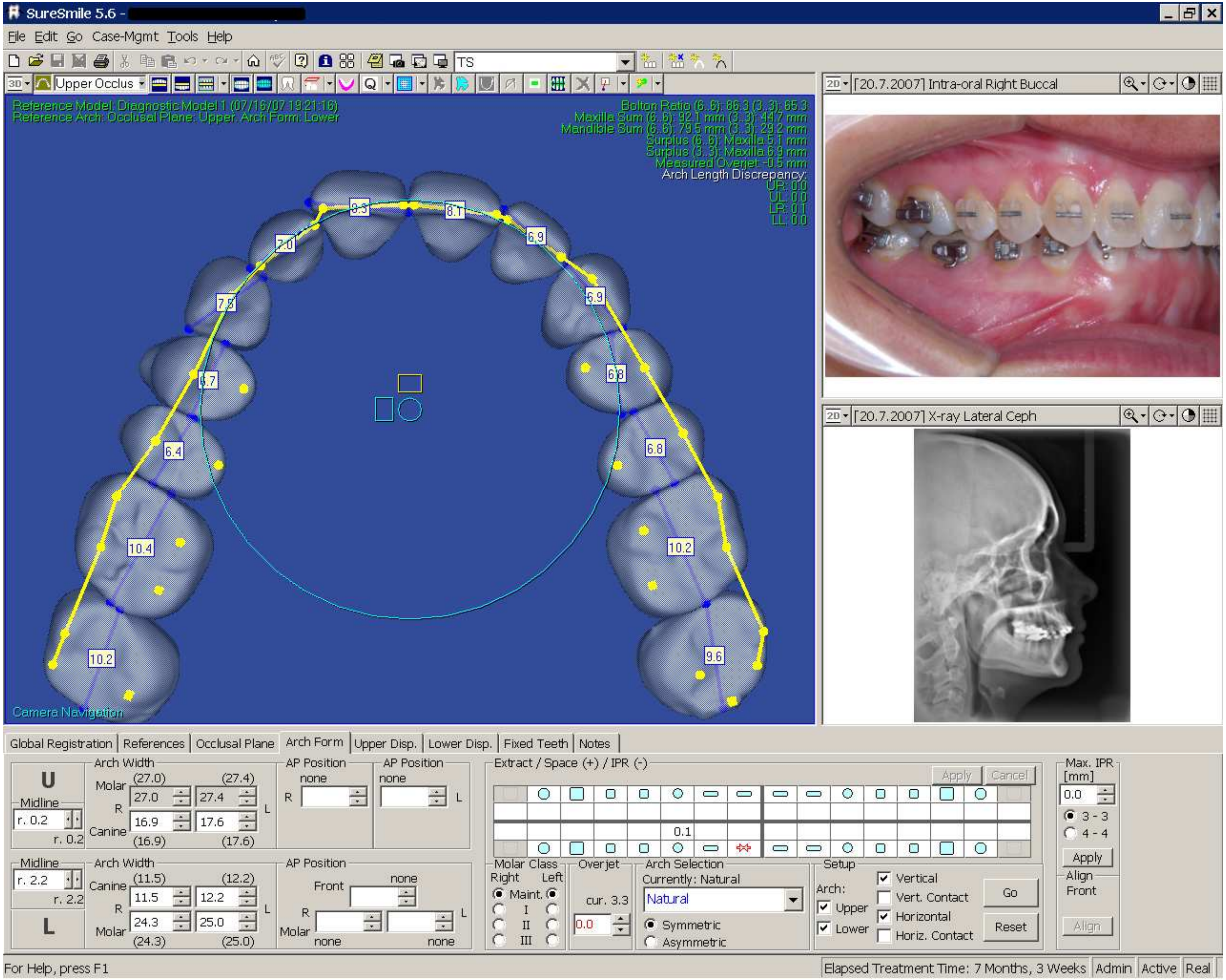
VRD Examples (Medicine)



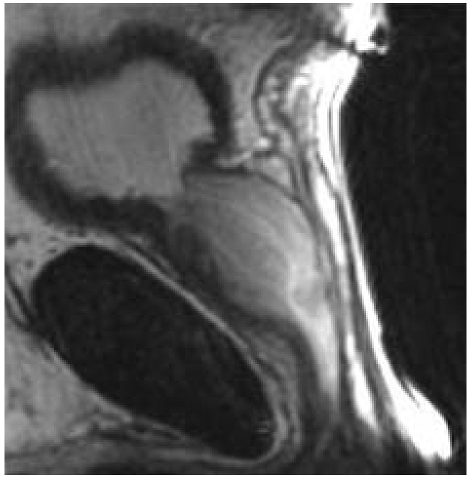
History of idea:
Medical applications

e.g. Cephalometry:

- A scientific study of the measurements of the head with relation to specific reference points
- utilizing a fixed, reproducible position for lateral radiographic exposure of skull
- used for orthodontic treatment planning, for evaluation of facial growth and development, including soft tissue profile.

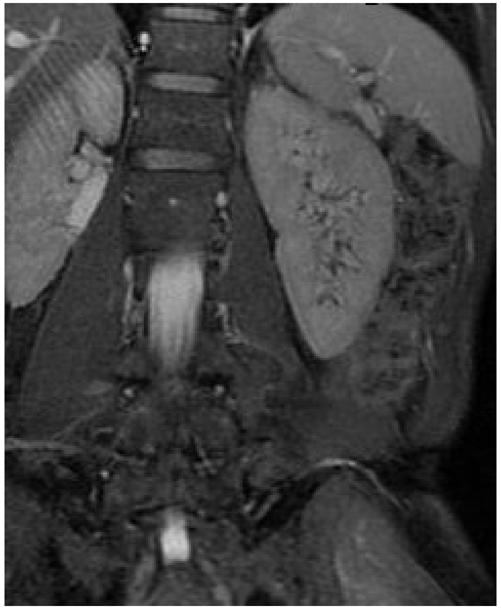

SureSmile 5.6 - [redacted]
File Edit Go Case-Mgmt Tools Help
TS
Upper Occlus
Reference Model: Diagnostic Model 1 (07/16/07 19:21:16)
Reference Arch: Occlusal Plane: Upper, Arch Form: Lower
Bolton Ratio (6..6): 86.3 (3..3): 65.3
Maxilla Sum (6..6): 92.1 mm (3..3): 44.7 mm
Mandible Sum (6..6): 79.5 mm (3..3): 29.2 mm
Surplus (6..6): Maxilla 5.1 mm
Surplus (3..3): Maxilla 6.9 mm
Measured Overjet -0.5 mm
Arch Length Discrepancy:
UR: 0.0
UL: 0.0
LR: 0.1
LL: 0.0
Camera Navigation
[20.7.2007] Intra-oral Right Buccal
[20.7.2007] X-ray Lateral Ceph
Global Registration
References
Occlusal Plane
Arch Form
Upper Disp.
Lower Disp.
Fixed Teeth
Notes
Arch Width
Molar
Canine
Midline
AP Position
none
Front
Extract / Space (+) / IPR (-)
Apply
Cancel
Max. IPR [mm]
3 - 3
4 - 4
Align Front
Align
Molar Class
Right
Left
Maint.
Overjet
cur. 3.3
Arch Selection
Currently: Natural
Natural
Symmetric
Asymmetric
Setup
Arch:
Upper
Lower
Vertical
Vert. Contact
Horizontal
Horiz. Contact
Go
Reset
For Help, press F1
Elapsed Treatment Time: 7 Months, 3 Weeks
Admin
Active
Real

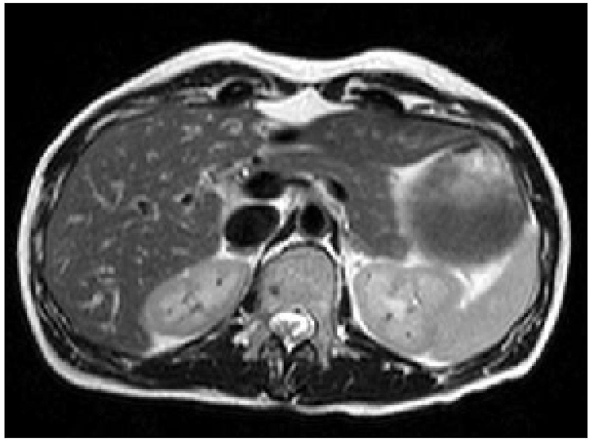
VRD Examples (Medicine)

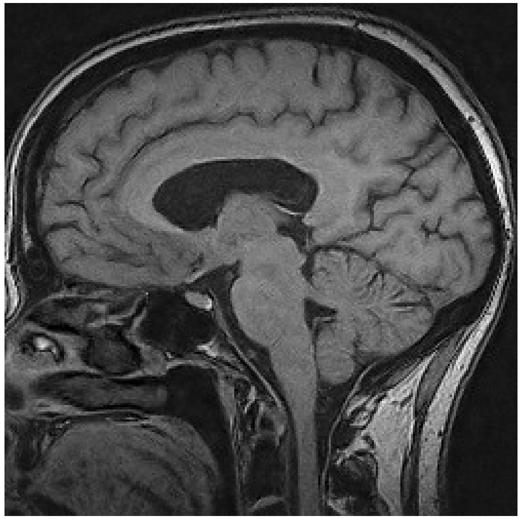


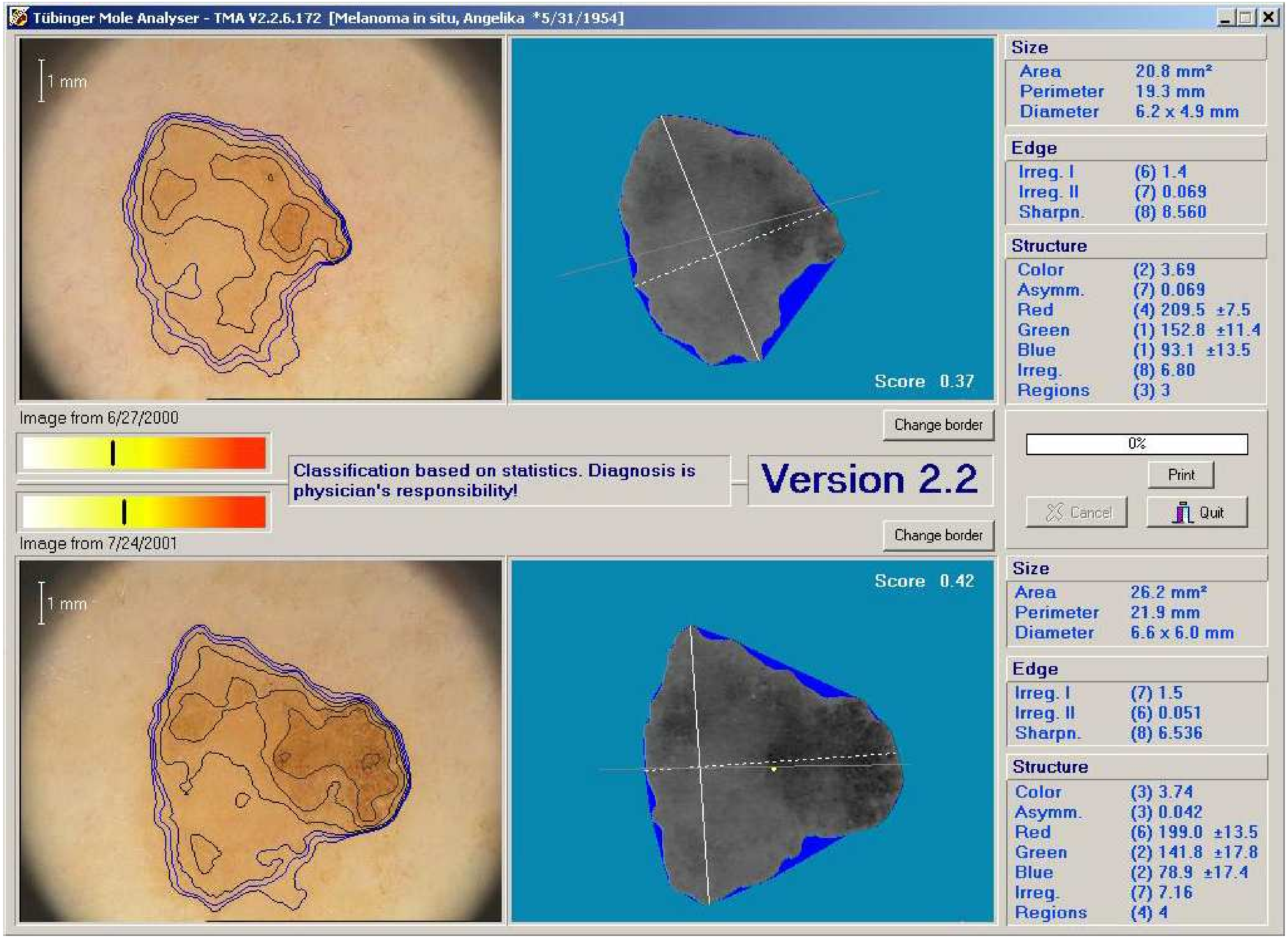
MRI prophylaxis

- Selection of frequent and serious diseases which are best detectable by MRI
- Description and quantification of decision relevant features (initially 2D, later 3D)
- Comparison with previous findings and cases





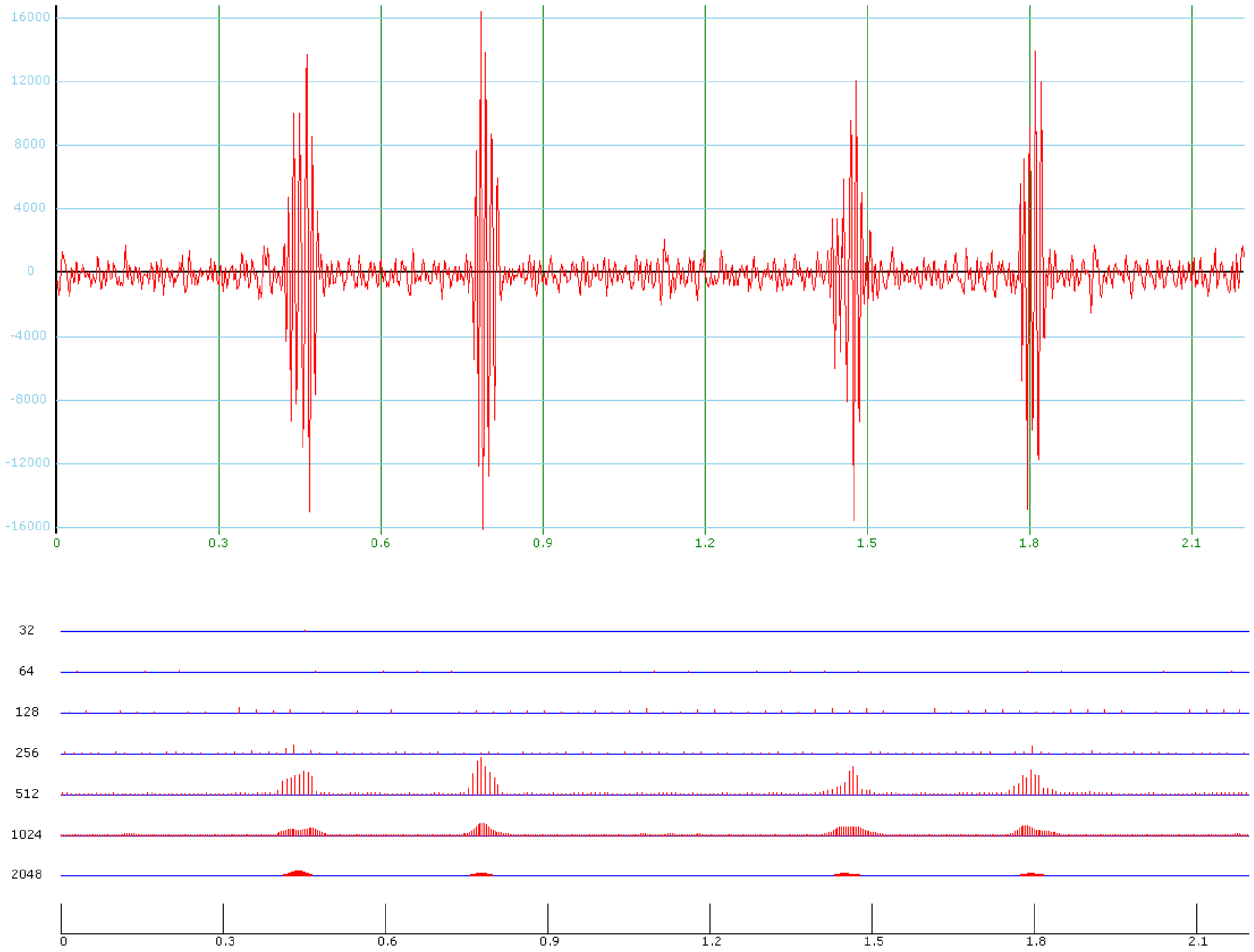




Source: Tübinger Moleanalyzer © FotoFinder Systems GmbH 2008

VRD Examples (Medicine)

Similarity search of Heart sounds



Home TestButton InsTestButton

◉ Pattern Search ○ Text Search ○ Login ○ Logout ○ Forgot Password ○ Change Password ○ New Account ○ Edit Account ○ Upload Publ ○ Config ○ Userlist ○ Options ○ Help

Here you can upload your pattern containing zip file to search for publications with similar patterns.

[ ] Durchsuchen...

Upload

Results to hs_shortnor.zip :

/p/1/2006_01_17_3/index.htm
title: shortnorm: a short normal heartsound
userid: author_userid_shortnorm
comment: example of comment
keywords: keyword_1_shortnorm; keyword_2_shortnorm; keyword_diagosis_shortnorm;
abs: text of abstract shortnorm ...
d: 0.0012175

/p/6/2006_01_26_1/index.htm
title: mitregur_6: a heartsound in case of mitral regurgitation (6)
userid: author_userid_mitregur_6
comment: example of comment mitregur_6
keywords: keyword_1_mitregur_6; keyword_2_mitregur_6; keyword_diagosis_mitregur_6;
sv: here optional text for searchview: searchviewtxt_mitregur_6
d: 0.5052

/p/1/2006_01_26_1/index.htm
title: mitregur: a heartsound in case of mitral regurgitation
userid: author_userid_mitregur
comment: example of comment mitregur
keywords: keyword_1_mitregur; keyword_2_mitregur; keyword_diagosis_mitregur;
sv: here optional text for searchview: searchviewtxt_mitregur
d: 0.5052

Example: Spinal surgery

More objectivity is possible:

1. Quantify decision relevant features
2. Compare feature vector worldwide
3. Decide for most successful treatment

Example for decision relevant features of osteoporotic compression fractures:

v = 2 * v2 / (v1+v3);

c = 4 * c2 / (v1+v3+d1+d3);

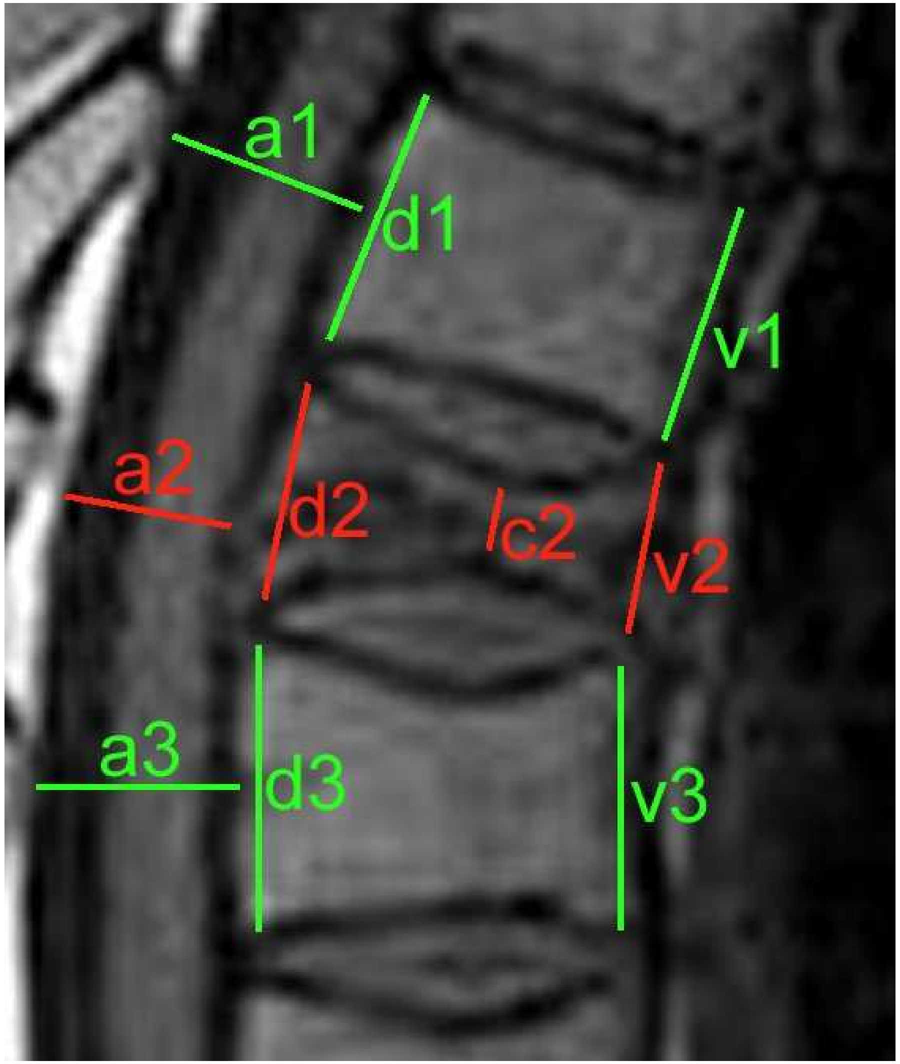
d = 2 * d2 / (d1+d3);

a = 2 * a2 / (a1+a3);

t = t-score (DXA bone density);

n = number of the vertebra;

Feature vector = (v, c, d, a, t, n);

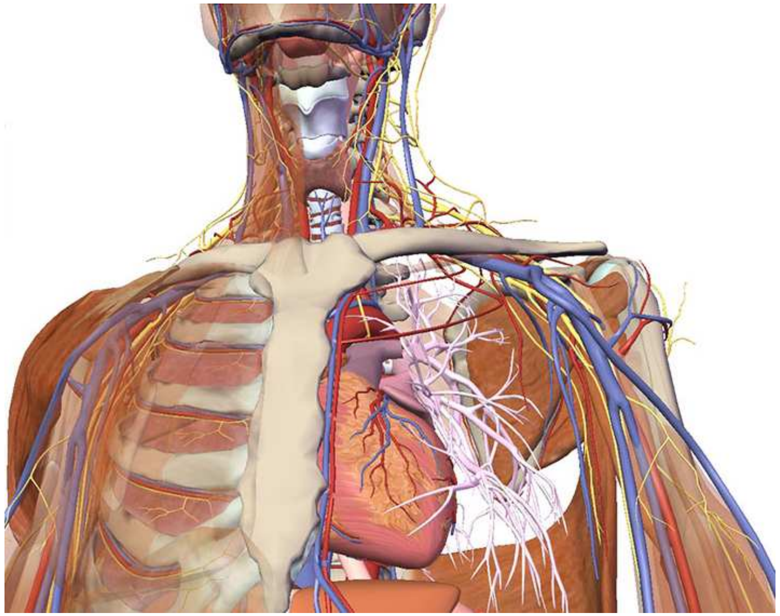


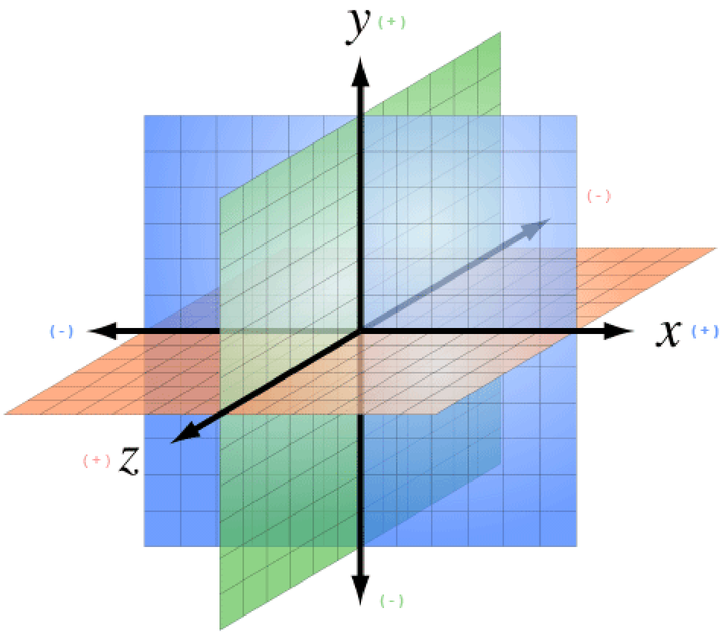
VRD Examples (Medicine)

Snomed extension: Anatomic reference model with 3D coordinate

System: Precise conversation possible using

3D Coordinates on reference model: Input of precise data





Source: http://www.visiblebody.com/Tour_Screen_shots

Next steps

Local implementation

- The next important step is an online prototype, which can be also expanded.
- Advanced implementations could support flexible queries and/or realize a first practical medical or commercial application.

Thank you for your attention!

Further Information: http://www.orthuber.com/wpa.htm

Email: orthuber@kfo-zmk.uni-kiel.de