Semi-automated contour recognition using DICOMautomaton

To cite this article: H Clark et al 2014 J. Phys.: Conf. Ser. 489 012088

View the article online for updates and enhancements.

Related content
- Automated segmentation and dose-volume analysis with DICOMautomaton
  H Clark, S Thomas, V Moiseenko et al.
- The Opportunity and Challenge of The Age of Big Data
  Hong Yunguo
- Shape descriptors for mode-shape recognition and model updating
  W Wang, J E Mottershead and C Mares
Semi-automated contour recognition using DICOMautomaton

H Clark1,2, J Wu2, V Moiseenko2,3, R Lee2, B Gill2, C Duzenli1,2 and S Thomas4

1 University of British Columbia, Vancouver, B.C., Canada.
2 British Columbia Cancer Agency - Vancouver Centre, Vancouver, B.C., Canada.
3 University of California San Diego, San Diego, Ca., USA.
4 British Columbia Cancer Agency - Fraser Valley Centre, Surrey, B.C., Canada.

E-mail: hdeanclark@gmail.com

Abstract. Purpose: A system has been developed which recognizes and classifies Digital Imaging and Communication in Medicine contour data with minimal human intervention. It allows researchers to overcome obstacles which tax analysis and mining systems, including inconsistent naming conventions and differences in data age or resolution.

Methods: Lexicographic and geometric analysis is used for recognition. Well-known lexicographic methods implemented include Levenshtein-Damerau, bag-of-characters, Double Metaphone, Soundex, and (word and character)-N-grams. Geometrical implementations include 3D Fourier Descriptors, probability spheres, boolean overlap, simple feature comparison (e.g. eccentricity, volume) and rule-based techniques. Both analyses implement custom, domain-specific modules (e.g. emphasis differentiating left/right organ variants). Contour labels from 60 head and neck patients are used for cross-validation.

Results: Mixed-lexicographical methods show an effective improvement in more than 10% of recognition attempts compared with a pure Levenshtein-Damerau approach when withholding 70% of the lexicon. Domain-specific and geometrical techniques further boost performance.

Conclusions: DICOMautomaton allows users to recognize contours semi-automatically. As usage increases and the lexicon is filled with additional structures, performance improves, increasing the overall utility of the system.

1. Purpose

The ever-growing volume of dosimetric patient data accumulated by medical centers is varied not only in terms of quality, scope, and format, but also in a more subtle way; each center has a unique naming and contouring dialect. While the Digital Imaging and Communication in Medicine (DICOM) standard addresses the former (i.e. interoperability between various types of hardware and software), it imposes constraints of the latter type on software. In particular, the flexibility of contour generation, specification, and labeling presents a challenge for systematic identification. For example, 60 patient contour sets from one centre over one year contained more than a dozen labels indicating the left parotid. In some cases both parotids were contoured into a single ‘parotids’ structure while in others the parotids had not been contoured - sometimes because they had been surgically removed. While an experienced researcher could reliably identify a left parotid, it is not conceivable to perform manual identification in analyses involving hundreds of patients. We present herein a set of practical lexicographical and geometrical...
techniques, incorporated into DICOMautomaton, for performing automated recognition of existing contour data.

2. Methods

We consider the situation whereby a researcher has a collection of existing contour data of mixed origin and a set of structures (e.g. organs) which are to be identified. An example might be locating all left parotids, taking into account that not every structure set is required to contain a left parotid and that unknown aliases may be used.

We define a lexicon \( l \) to be the collection of exact mappings from a set of raw labels \( d_i \) to a unique label \( c_i \) such that \( l(s) = c_i \) if and only if \( s \) is a label in \( d_i \). We denote the set of all \( d_i \) as \( d \) and the set of all \( c_i \) as \( c \). The lexicon represents the researcher’s domain knowledge and is similar to a thesaurus. For instance a limited, toy lexicon for head and neck cancers might be

\[
\begin{align*}
d_1 &= \{ \text{parotid}, \text{lt., par, Left Parotid} \} \\
d_2 &= \{ \text{rt. partd., r. par, Right Parotid} \} \\
d_3 &= \{ \text{chiasm, opt, chiasm} \}
\end{align*}
\]

Using this lexicon, one would be unable to identify a structure named “left par” though it is clear that it most closely relates to \( d_1 \). To capture this intuitive notion of similarity we model group membership using string similarity measures, which consider two inputs similar if they share specific features. The similarity between strings \( s_1 \) and \( s_2 \) for measure \( J \) is written as a score \( S_J(s_1, s_2) \in [0,1] \) where \( S_J = 0 \) denotes no similarity and \( S_J = 1 \) perfect similarity. Scoring allows us to incorporate the similarity of strings from many measures. Given a lexicon \( l \) and input \( s \) which refers to an unknown member of \( c \), we choose a combining function \( F \) to weight the similarity scores with the elements of \( d_i \) into a total score for \( c_i \)

\[
T_{c_i}(s) = F \left[ \left\{ S_J(s, \hat{s}) \mid \hat{s} \in d_i \right\} ; l \right] \in [0,1].
\]

Typically, \( F \) is chosen to help reduce statistical uncertainty by producing a high \( T_{c_i}(s) \) when many measures produce high scores and a low \( T_{c_i}(s) \) otherwise. For well behaved general measures, \( F \) could be a simple mean. Denoting the set of all \( T_{c_i}(s) \) as \( T(s) \), the suspected translation \( \hat{c} \) is determined by choosing a mapping function \( M(T(s)) = \hat{c} \mid \hat{c} \in c \). Both \( F \) and \( M \) are subject to the desirable property that if \( s \in d_i \) then \( \hat{c} = c_i \) as one would find using the lexicon without any consideration of similarity. We can judge the validity of \( \hat{c} \) using \( T(s) \).

Due to wide diversity of contour labeling encountered, \( F \) and \( M \) were equipped with tunable minimum thresholds to help reliably control type I and II errors. For this study, \( F \) was implemented as a weighted-average while \( M \) was taken to be a maximum function with additional degeneracy-handling logic.

Well-known lexicographic measures implemented in DICOMautomaton include Levenshtein-Damerau [1] and Jaro-Winkler [2] edit measures, Soundex [3], Double Metaphone [4], and Match Rating Approach [5] phonetic measures, statistical measures including longest common substrings/sequences, N-grams of user-specified order \( N \) (character or word based in a variety of flavours, including Dice [6], Jaccard [7], and cosine metrics), and a generic bag-of-characters measure. The Levenshtein-Damerau measure is popular due to its speed and simplicity. Alternative lexicographical measures implemented include an artificial neural network-based measure, a self-orthogonalizing measure which ignores all common N-grams by elements of \( d \), a custom measure (which we refer to as DICOMhash) which differentiates labels which are liable to confuse other techniques (e.g. ‘CTV+3mm’ vs. ‘CTV+5mm’), and a domain-specific head and neck measure which hard-codes an individual centre’s naming conventions.

Geometrical methods considered are similar to the aforementioned lexicographic methods; instead of a string input, they take a set of ordered two-dimensional contours. They rely on
geometrical features, but otherwise serve an identical purpose as the lexicographic measures. Geometric measures implemented include probability spheres and boolean overlap relative-position measures, Fourier descriptor and eccentricity shape-based measures, a domain-specific measure, and simple feature measures involving volume, total perimeter/surface area, and centroids [8]. The most intuitive measures involve the spatial location of objects in $\mathbb{R}^3$, so we describe here probability spheres to demonstrate the basic approach: in normalized coordinates, each target structure is given a centre, an effective radius, and a radially-dependent normalized weighting $W(\vec{r})$. The purpose of weighting is to prescribe inhomogeneous regional similarity. The similarity score between structures is evaluated as $\int W_i(\vec{r})W_2(\vec{r})d\vec{r}$. The more structures overlap in $\mathbb{R}^3$, the more likely we are to think of them as similar.

Cross-Validation Lexicon Folding (cvlf) was used to estimate the overall effectiveness of the system. cvlf involves randomly choosing a fraction $f$ of $d$ and measuring the ability of the system to correctly translate the entirety of the complete lexicon. This is loosely analogous to measuring the abilities of an individual to reconstruct a foreign vocabulary after being given only a sample portion. Random selection of a portion of the lexicon may result in omission of $c_i$, yielding a system artificially unable to recognize any of the corresponding $d_i$. Where applicable we correct for this effect. Denoting the number of elements in $d$ as $N_d$, omission of $f N_d$ elements reduces the maximum cvlf recognition rate, on average, to $1 - \sum_i P(c_i) N_d / N_d$ where

$$P(c_i) = \frac{\Gamma(N_d - N_{d_i} + 1) \Gamma(N_d - f N_{d_i} + 1)}{\Gamma(N_d - (f + 1) N_{d_i} + 1) \Gamma(N_d + 1)} \bigg|_{f N_d \leq N_d - N_{d_i}}, \text{otherwise } 0.$$ 

To evaluate system performance, contour label data from 60 head and neck cancer patients was used for cvlf, producing a lexicon composed of $N_d = 325$ and $N_c = 18$ elements denoting 16 unique structures and two honey-pots (i.e. for interception of artifacts). The honey-pots comprised 146 strings and the remaining were distributed with an average $N_d$ of 11.2 ($\sigma = 5.1$).

3. Results

$M$ and $F$ thresholds were found to greatly affect the recognition and error rates. Both recognition and type $I$ error rates increased when decreasing thresholds. Type $II$ error rates increased when increasing thresholds. Reasonable default values were found to be 0.3 for $F$ and 0.5 for $M$.

Results of cvlf for various retention fractions $f$ are shown in figure[1]. We estimate that in typical circumstances a user would possess between 50 – 80% of $d$ and 75 – 100% of $c$ from the complete lexicon. Focusing on an information-deficient situation where $f = 0.3$ (i.e. 30% of $d$ are known), the Levenshtein-Damerau measure performed successful recognition 76% of the time with type $I$ errors occurring 23% of the time. The use of several (mixed) lexicographic measures improved the raw recognition rate 7% and reduced type $I$ errors 6%, representing an effective overall improvement in more than 10% of recognition attempts.

Geometrical techniques performed similar to lexicographic techniques. In general, they produced both lower recognition and error rates compared with lexicographic techniques. Remarkably, the type $I$ error rate did not increase above 7% at any $f$ using mixed geometrical measures (probability spheres, perimeter length, a lateral-position discriminator, and centroid comparison). At $f = 0.3$, mixed geometrical measures performed successful recognition 71% of the time. For probability spheres, it was found that an appropriate effective radii was $r_{eff} = 3 \left(3V/4\pi\right)^{1/3}$ where $V$ was the structure’s volume. The preceding factor of 3 was chosen to help handle variations in position due to patient geometry and orientation.

Domain-specific lexicographic measures which hard-coded naming conventions performed dramatically better, achieving an optimal recognition rate of 97% when $f = 0.3$. 
Figure 1. Comparison of select measures. Recognition (left) and type I error (right) rates versus lexicon retention during cvlf. Also shown are the optimal recognition and exact match rates. The latter refers to application of $l$ without the use of similarity measures. The mixed lexicographical measure contains Levenshtein-Damerau, Jaro-Winkler, DICOMhash, N-gram, and bag-of-characters measures. The mixed geometrical measure contains probability spheres, perimeter length, a lateral-position discriminator, and centroid comparison measures. The domain-specific (lexicographic) measure achieves optimal recognition.

4. Discussion
The use of mixed lexicographic measures resulted in an effective overall improvement in more than 10% of recognition attempts compared with the Levenshtein-Damerau approach.

Although geometrical techniques generally showed reduced error rates, they required considerably more processing power, memory, and storage to use. Furthermore, they introduced subtle dependencies on data fidelity and were somewhat unwieldy compared to simpler lexicographic techniques.

Domain-specific measures performed optimally in recognition. Unfortunately, they are by nature often incapable of handling previously-unseen input. It is unclear how to appropriately gauge their performance during real usage. A balance between the uncertainty of domain-specific measures and the less-performant, but more robust, lexicographic and geometrical measures can be achieved with appropriate mixing. However an estimate of $f$ is required for reliable weighting, which may be difficult to assess.

Hybrid techniques could further improve system efficiency by involving meta-information. For example, simultaneous recognition on a collection of mutually exclusive input can ensure that the system does not erroneously detect two spinal cords in a single patient.

Actual usage indicates performance higher than that suggested by cvlf. We believe this is
due to CVLF randomly removing elements of \( d \); in reality some labels are encountered frequently (e.g. ‘body’) while others are rare. This regularity provides increased recognition power.

Finally, we hope to have demonstrated that recognition is, in some cases, a flexible way to handle contour data. It rejects the rigidly-defined database paradigm and encourages data sharing across labeling and contouring conventions.

5. Conclusions

DICOMautomaton can be used to perform accurate, semi-autonomous contour recognition. Lexicographic methods are generally suitable to this end when the domain is well-known, while more computationally-burdensome geometrical methods are available for information-deficient situations. Mixing lexicographic measures produces an effective recognition improvement of more than 10% over a pure-Levenshtein-Damerau approach, while domain-specific measures can achieve optimal recognition. Increasing domain knowledge in the lexicon increases performance, and so continued usage will tend to increase the successful recognition rate.

Acknowledgments

The authors acknowledge financial support from The Ride to Conquer Cancer and the BCCA.

References

[1] Levenshtein V 1966 Binary codes capable of correcting deletions, insertions and reversals *Soviet Physics Doklady* vol 10 p 707

[2] Winkler W E 1990 String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage *Survey Research Methods* (American Statistical Association) pp 354–359

[3] Knuth D E 1973 *The art of computer programming, Vol. 3* (Addison-Wesley, Reading, MA)

[4] Philips L 2000 *CC Plus Plus Users Journal* 18 38–43

[5] Moore G B, Kuhns J L, Trefftzs J L and Montgomery C A 1977 *Accessing individual records from personal data files using non-unique identifiers* (US Department of Commerce, National Bureau of Standards)

[6] Dice L R 1945 *Ecology* 26 297–302

[7] Jaccard P 1901 *Etude comparative de la distribution florale dans une portion des Alpes et du Jura* (Impr. Corbaz)

[8] Nixon M and Aguado A S 2008 *Feature extraction & image processing* (Academic Press)