Database Article

Isfahan MISP Dataset

Abstract

An online depository was introduced to share clinical ground truth with the public and provide open access for researchers to evaluate their computer-aided algorithms. PHP was used for web programming and MySQL for database managing. The website was entitled "biosigdata.com." It was a fast, secure, and easy-to-use online database for medical signals and images. Freely registered users could download the datasets and could also share their own supplementary materials while maintaining their privacies (citation and fee). Commenting was also available for all datasets, and automatic sitemap and semi-automatic SEO indexing have been set for the site. A comprehensive list of available websites for medical datasets is also presented as a Supplementary.

Keywords: *Algorithms, computers, dataset, online*

Masoud Kashefpur, Rahele Kafieh[1], Sahar Jorjandi, Hadis Golmohammadi, Zahra Khodabande, Mohammadreza Abbasi, Nilufar Teifuri, Ali Akbar Fakharzadeh[2], Maryam Kashefpoor[3], Hossein Rabbani[1]

Student Research Committee, [1]Medical Image and Signal Processing Research Center, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran, [2]Department of Computer Engineering, Higher Education Institute of Allame Naeini, Naein, Isfahan, Iran, [3]Institute for Cognitive and Brain Sciences, Shahid Beheshti University, Tehran, Iran

Introduction

To stimulate the advancement of computer-aided diagnostic research for medical image and signals, there is need to have an online depository to share clinical ground truth with the public and provide open access for researchers to evaluate their computer-aided algorithms.[1-4] Such databases are collected laboriously and are not commonly shared with the community. New methods and systems on medical image and signals are mostly based on datasets by each developer; however, such a strategy makes the performance comparison imperfect and unreliable. The perfect performance testing demands a solid evaluation using "case data" from different modalities and after approval by medical experts.[4] To clear up the word "case data," an example on mammography could be helpful; in such a case, at least two views of each breast with digital radiography should be provided along with magnetic resonance imaging (MRI) or ultrasound.[4] Furthermore, the data classification (to normal and disease) and possibly regions of abnormality labeled by the expert as "ground truth" or "gold standard" or "interpretation" are needed for a perfect "case data." The main idea behind designing this website was to provide a collection concentrated on medical datasets rather than only a general data collection.

Comprehensive collection of available online datasets for medical image and signal processing (MISP)

There were only a few attempts made to establish appropriate reference datasets for medical imaging and signal processing[2] until the past 10 years. The available datasets did not include perfect "case data" because most of them had no ground truth and almost all of them were not free of charge. However, undeniable need for such datasets led to numerous online medical image and signal databases in recent years.

There are also some websites that list and/or host multiple collections of data. Table 1 is a list of such websites. Furthermore, we provide a full list of databases (in mentioned websites or other individual sources) as a Supplementary.

In this study, we introduce Isfahan MISP dataset website and present a short description of each dataset category. Furthermore, we introduce a collection of available datasets (sorted according to anatomical/physiological information and imaging modality), which are freely available on the internet. There are also some websites that list and/or host multiple collections of data, and we shortly discuss about them.

How to cite this article: Kashefpur M, Kafieh R, Jorjandi S, Golmohammadi H, Khodabande Z, Abbasi M, *et al.* Isfahan MISP dataset. J Med Sign Sence 2017;7:43-8.

Address for correspondence: *Dr. Rahele Kafieh, Medical Image and Signal Processing Research Center, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran. E-mail: rkafieh@gmail.com*

Website: www.jmss.mui.ac.ir

Table 1: Some websites that list and/or host multiple collections of medical datasets

CVonline: Image Databases:	http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm
DATABASES ON MEDICINE AND MOLECULAR BIOLOGY	http://www.meddb.info/index.php.en?cat=1
Database of cancer MR,CT,PET, . . . from lung, brain	http://www.cancerimagingarchive.net/
National Biomedical Imaging Archive (NBIA):	https://imaging.nci.nih.gov/ncia/login.jsf
• Lung Image Database Consortium (LIDC)	
• Reference Image Database to Evaluate Response (RIDER)	
• Breast MRI	
• Lung PET/CT	
• Neuro MRI	
• CT Colongraphy	
• Virtual Colonoscopy	
• Osteoarthritis Initiative	
• PET/CT phantom scan collection	
BIRN / XNat	http://www.birncommunity.org/resources/data/
• High SNR Healthy Volunteer DTI Calibration Dataset	
• Morphometry BIRN Multi-site Multi-session Structural MRI Data	
• Open Access Structural Imaging Series (OASIS)	
• Duke Center for In-Vivo Microscopy High Resolution MRI Images	
MIDAS	http://www.insight-journal.org/midas/
• National Alliance for Medical Image Computing (NAMIC)	
○ Lupus white matter lesions	
○ Brain MRI: 2-4 years old	
○ Prostate	
• NLM: Imaging Methods Assessment and Reporting	
• Liver tumors with segmentations	
100 Healthy Brain MRI: 18-90 years old	http://www.insight-journal.org/midas/community/view/21
UCI Machine Learning Repository: The father of internet data achives for all forms of machine learning.	http://archive.ics.uci.edu/ml/
Cornell Visualization and Image Analysis (VIA) group: Provides a list of available databases, many of which are also listed here.	http://www.via.cornell.edu/databases/
UT Health Scient Center Image Collections: List of medical images, atlases, and databases available on the web.	http://www.library.uthscsa.edu/find/databases.cfm?Category=Image%20Collections
OmniMedicalSearch.com: Medical Image Databases & Libraries	http://www.omnimedicalsearch.com/image_databases.html

Construction and Content

The working group in MISP research center had set one of its goals to develop a reference medical image and signal database for image and signal processing research and development. The main aim of this collection was to provide reliable datasets for researchers to compare their algorithms on common framework and to be able to interpret and compare the performance of their newly developed methods. The collected databases were devised to be in accordance with patient privacy and copyright issues. They all have written and signed forms by the institutional review board, also known as an independent ethics committee. The ground truth was provided for datasets based on evaluation of one, two, or three experts. Some datasets were collected from different modalities or from diverse anatomical structures of one patient. Most of the mentioned datasets were available with a published paper in famous journals to demonstrate possible application of the dataset.

Technical background

In this project, MySQL was used as the backend database. MySQL is an open-source database management system. The features of MySQL are listed below:

- *Relational database management system*. A relational database stores information in different tables, rather than in one giant table. These tables can be referenced to each other, to access and maintain data easily.

- *Open source database system*. The database software can be used and modified by anyone according to their needs.
- *Fast, reliable, and easy-to-use*. MySQL is multithreaded database engine, and this feature improves its performance. A multithreaded application performs many tasks at the same time as if multiple instances of that application were running simultaneously.

Multithreaded MySQL has many advantages. A separate thread (which is always running) handles each incoming connection and manages the connections. Multiple clients can perform "read" operations simultaneously, but while writing, only one client, who needs access to updated data, is held up. Even though the threads share the same process space, they execute individually, and because of this separation, multiprocessor machines can spread the thread across many central processing units (CPUs) as long as the host operating system (OS) supports multiple CPUs. Multithreading is the key feature to support MySQL's performance design goals, and it can be counted as the core feature around which MySQL is built.

The proposed web programming for this website is PHP. PHP is server side scripting language compatible with all known OSs like Windows, Linux, etc. Nowadays, PHP frameworks are being extensively used by programmers to address the performance tuning issues faster and with ease. They offer extensible architecture and features that make source code programming easier by providing standard templates and plug-ins. The main features of PHP are listed below:

- *Powerful database-driven functionality*: PHP integrates well with MySQL and contains a library full of useful functions to help the usage of MySQL. There are even many database managers written in PHP.
- *Faster web applications*: Because PHP does not use a lot of a system's resources to run, it operates much faster than other scripting languages. Hosting PHP is also very easy and lots of hosts provide support for PHP. Even when used with other softwares, PHP still retains speed without slowing down other processes. It can be concluded that PHP is a mature language, and it is also fairly stable because all the kinks have been worked out over the years.
- *Object oriented*: PHP actually has the ability to call Java and Windows COM objects. In addition to this, one can create custom classes. Other classes can actually borrow from those custom classes and this specification extends the capabilities of PHP even further.
- *High level freedom*: When comparing PHP to a language such as ASPX, the level of freedom is far superior. As mentioned before, PHP is open-source. One can use any text editor to code PHP such as Notebook++, jEdit, Emacs, Bluefish, or even just Notepad if felt inclined. In cases like development of applications with ASPX, Microsoft Visual Studio will be a limited solution. PHP also is not OS specific and can be run on all well-known OS's.
- *Free of charge*: There are no costs associated with using PHP, including updates. Keeping costs down is a goal of any business and developers as well. Therefore, the fact that one can code programs with PHP for free is a huge benefit that will not be provided with JPS, ASP, or other scripting languages that require paid hosting. There are no licenses, restrictions, or royalty fees involved at all, and PHP is 100% free for anyone to use.

Regarding above features, we used PHP for web programming and MySQL for database managing. The resulted "biosigdata.com" was a fast, secure, and easy-to-use online database for medical signals and images. Among these features, "biosigdata.com" provided registered users the ability to download medical signals and images selectively and freely. It was also an appropriate place for researchers around the world to share their data and supplementary materials with each others, because users could freely, easily, securely, and unlimitedly upload their datasets while managing their privacies (citation and fee). Uploaded datasets would be categorized and published after verification. Commenting and replying for the comments were available for all datasets; therefore, researchers could share their experiences. To improve functionality in search engines (e.g., Google), automatic sitemap and semi-automatic SEO indexing have been set for the site.

Sample description for available sets in MISP dataset website

To provide more description on contents of MISP dataset, two examples of the available sets are described below. As a long-term goal of MISP research center, the datasets would be added to this website, and it was open to accept databases from other researchers by convincing them to the reality that making online free datasets can improve the research quality of the owner and the users.

The first sample dataset was on ocular images and included the following image sets:

(1) MISP Isfahan optical coherence tomography (OCT) dataset for segmentation.[5]
This dataset contained 13 three-dimensional (3D) macular SD-OCT images obtained from eyes without pathologies using Topcon 3D OCT-1000 imaging system in Ophthalmology Dept., Feiz Hospital, Isfahan, Iran. The datasets were in mat format and were named "1" to "13." The x, y, z size of the obtained volumes was $512 \times 650 \times 128$ voxels, $7 \times 3.125 \times 3.125$ mm^3, and voxel size $13.67 \times 4.81 \times 24.41$ μm^3 [as shown in Figure 1]. The values of the proposed layer localization were provided for 12 boundaries and were stored in another mat file named "randomImages" (the rows corresponded to each dataset and the columns contained the traced

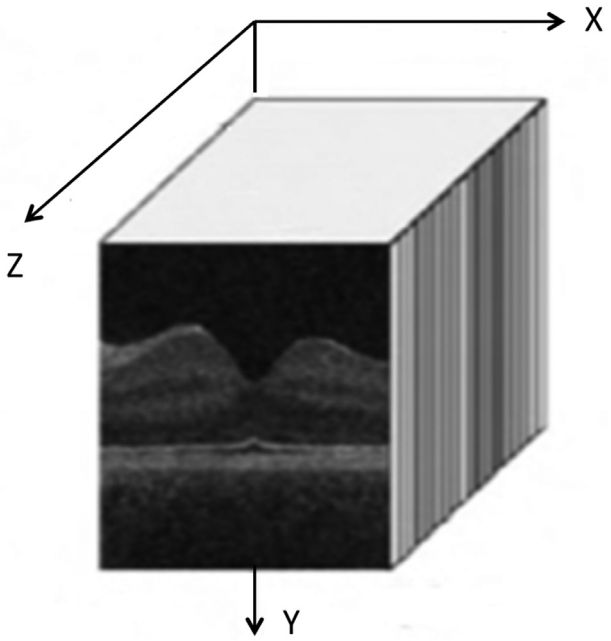

Figure 1: The x, y, z size of the obtained volumes was 512 × 650 × 128 voxels[5]

slices). Furthermore, the validation was based on manual tracing by two observers, and mean value of the mentioned tracings were provided in mat files named "totalManual_1" to "totalManual_13."

(2) MISP Isfahan OCT dataset for denoising.[6,7]
This dataset contained six 3D OCT data using Topcon 3D OCT-1000 imaging system in Ophthalmology Dept., Feiz Hospital, Isfahan, Iran. The datasets were in mat format and were named "1" to "6." Participants in the dataset were diagnosed to have retinal pigment epithelial detachment. The positions of the mentioned slices were stored in another mat file named "random_Topcon."

(3) MISP Isfahan OCT dataset for vessel segmentation of OCT and Scanning Laser Ophthalmology (SLO).[8]
This dataset contained 24 3D macular SD-OCT images (along with SLO images) obtained from eyes without pathologies using Heidelberg HRA OCT scanner imaging system in Ophthalmology Dept., Noor Hospital, Tehran, Iran. Each dataset consisted of a SLO image and limited number of OCT scans with size of 496 × 512 (namely, for a data with 19 selected OCT slices, the whole data size was 496 × 512 × 19). All pixels in the SLO images were labeled as either "vessel" or "non-vessel" manually to make a standard in evaluation of the proposed method. The datasets were in mat format and were named "1" to "24."

(4) MISP Isfahan OCT dataset from normal population.[9]
This dataset contained 112 3D macular SD-OCT images (along with SLO images) obtained from eyes without pathologies from 76 people using Heidelberg HRA OCT scanner imaging system in Ophthalmology Dept., Noor Hospital, Tehran, Iran. Each dataset consisted of a SLO image and limited number of OCT scans with size of 496 × 512. The datasets were in mat format and were named "1" to "112."

(5) MISP Isfahan enhance depth imaging OCT dataset from normal population.[10]
This dataset contained 19 3D macular SD-OCT images (along with SLO images) obtained using Heidelberg HRA OCT scanner imaging system in Ophthalmology Dept., Noor Hospital, Tehran, Iran. Each dataset consisted of a SLO image and limited number of OCT scans with size of 496 × 512. The datasets were in mat format and were named "1" to "19."

(6) MISP Isfahan corneal OCT dataset from normal population.[11,12]
This dataset contained 15 3D corneal SD-OCT images obtained using Heidelberg HRA OCT scanner imaging system in Ophthalmology Dept., Noor Hospital, Tehran, Iran. Each dataset consisted of a limited number of OCT scans with size of 496 × 512. The datasets were in mat format and were named "1" to "15."

The second sample dataset was on audio and acoustic signals and included the following datasets:

(1) Voice samples of patients with Parkinson's disease (spontaneous swallows in Parkinson's disease).[13]
Data were collected from 34 patients (19 males) who suffered from Parkinson's disease (PD) (age = 59.85 ± 11.46 years). They also had videofluoroscopy swallow study assessment at the same day. Patients' demographic data, including age, sex, time from commencement of disease, and the Hoehen and Yahr scale score, which shows the stage of PD, were collected via an informed interview. All participants or their legal guardian signed a consent form to participate in the experiments. The study had been approved by the Biomedical Ethics Board of Isfahan University of Medical Sciences. Each patient participated in the following two tests: sound recording and oropharyngeal videofluoroscopy. The two tests were performed independently, whereas videofluoroscopy study was done to validate which participant suffered from aspiration generally.

Acoustic recording study: The acoustic recording was conducted in the test room with indoor lighting, which was at 24°C. There was ambient noise from outside of the room but it was minimized during the recordings. Swallowing sounds were recorded via microphone (C417 omnidirectional condenser lavalier microphone, AKG Acoustics, Austria) connected t o a portable digital sound recorder (Edirol R-44, Japan). The microphone was fixed at the center of a 1.5-cm-diameter rubber piece. This rubber piece was attached to the surface of the skin using double-sided

adhesive tape. The recorder amplified and digitized the received signal at a sampling rate of 44.1 kHz. The microphone was positioned over the laryngopharynx of each patient for 15 min when the participant was in a seated position watching a nature movie. Because no food was used in this test, spontaneous swallows that were registered were innate saliva swallows. The number of swallows that each participant had within the 15-min time period was varied. All data recordings were carried out by a speech language pathologist and a biomedical engineer.

(2) Voice samples of patients with internal nasal valve collapse before and after functional rhinoplasty.[14]

This dataset contained voice recordings of participants who had internal nasal valve collapse. These voice samples were corresponding to the following two groups: before and after functional rhinoplasty. The initial voice samples were recorded 1–15 days before the surgical operation. Inclusion criteria were the lack of cold, hoarseness in all participants, and menstruation in females on the voice recording day. A demographic questionnaire was completed by the participants. The participants were familiarized with the method of the test before recordings. To minimize the environmental noise, recordings were done in an acoustic room (with noise of 28 dB based on sound level meter [TES-1351]). The voice samples were recorded separately for each participant. The participants were in a seated position and a microphone (Somic SENIC ST-818 3.5 mm on-ear stereo headphones with adjustable microphone and 2.5 m cable) was located 5 cm from the center of their lips on the right corner of mouth.

The participants were guided to generate the vowels /a/ and /i/ in three tests for a period of 5 s. The reason for choosing these vowels was their different location of articulation in the vocal tract. /i/ is the highest front vowel, and /a/ is the lowest back vowel.

Vocal intensity during vowel prolongation was held at 75 ± 2 dB (measured with sound level meter) by asking each participant in person. The patients' voice sample was recorded at the same time using Praat (version 5.0.23) software[15] at a sampling frequency of 44.1 kHz using a laptop equipped with a sound card. The surgical operation in all patients was performed by the same surgery team. Voice signals were obtained again 3 months after the surgery under the same conditions. Prior to recording the second voice samples, the participants underwent examinations by an ENT specialist and also were examined again in terms of inclusion and exclusion criteria of the study. The collected voice samples were encoded by a person who was not involved in the analysis of voice signals.

Conclusion

The proposed online dataset aims at filling the gap in medical image and signal depository, which can provide the sets to be downloaded and uploaded with different privacies set by the owners. The "biosigdata.com" is a free service, and the structure makes it a fast, secure, and easy-to-use online database.

Acknowledgements

The authors thank the following people (arranged alphabetically) for gathering parts of datasets in the website or for providing web addresses for databases: Dr. Mohammadreza Akhlagi, Dr. Zahra Amini, Dr. Majid Barekatein, Dr. Nasim Dadashi, Hajar Danesh, Dr. Mahdad Esmaili, Iman Fartah Abadi, Amin Farah Abadi, Maria Farahi, Marzieh Golabbakhsh, Shirimn Hajeb, Dr. Fedra Hajizade, Jalil Jalili, Mehdi Kazemian, Dr. Saeed Kermani, Tahere Mahmoodi, Dr. Alireza Mehri, Mohammadreza Momenzadeh, Dr. Alireza Peyman, Atie Purfarmanbar, Farhad Rahimi, Zahra Saeedi Zadeh, Nilufar Salehi, Dr. Omid Sarrafzadeh, Dr. Mohammadreza Sehhati, Ramin Soltanzadeh, Dr. Ardeshir Talebi, Dr. Alireza Vard, and Hossein Yusefi.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Horsch A, Hapfelmeier A, Elter M. Needs assessment for next generation computer-aided mammography reference image databases and evaluation studies. Int J Comput Assist Radiol Surg 2011;6:749-67.
2. Horsch A, Prinz M, Schneider S, Sipilä O, Spinnler K, Vallée J-P, *et al.* Establishing an international reference image database for research and development in medical image processing. Bildverarb Med 2003;363-7.
3. Horsch A, Blank R, Eigenmann D, editors. EFMI Reference Image Database Initiative: Concept, State and Related Work. International Congress Series; 2005: Elsevier.
4. Deserno TM, Welter P, Horsch A. Towards a repository for standardized medical image and signal case data annotated with ground truth. J Digit Imaging 2012;25:213-26.
5. Kafieh R, Rabbani H, Abramoff MD, Sonka M. Intra-retinal layer segmentation of 3D optical coherence tomography using coarse grained diffusion map. Med Image Anal 2013;17:907-28.
6. Kafieh R, Rabbani H, Abramoff MD, Sonka M. Curvature correction of retinal OCTs using graph-based geometry detection. Phys Med Biol 2013;58:2925-38.
7. Kafieh R, Rabbani H, Selesnick I. Three dimensional data-driven multi scale atomic representation of optical coherence tomography. IEEE Trans Med Imaging 2015;34:1042-62.

8. Kafieh R, Rabbani H, Hajizadeh F, Ommani M. An accurate multimodal 3D vessel segmentation method based on brightness variations on OCT layers and curvelet domain fundus image analysis. IEEE Trans Biomed Eng 2013;60: 2815-23.
9. Kafieh R, Rabbani H, Hajizadeh F, Abramoff MD, Sonka M. Thickness mapping of eleven retinal layers segmented using the diffusion maps method in normal eyes. J Ophthalmol 2015;2015: 259123.
10. Danesh H, Kafieh R, Rabbani H, Hajizadeh F. Segmentation of choroidal boundary in enhanced depth imaging OCTs using a multiresolution texture based modeling in graph cuts. Comput Math Methods Med 2014;2014:479268.
11. Jahromi MK, Kafieh R, Rabbani H, Dehnavi AM, Peyman A, Hajizadeh F, *et al.* An automatic algorithm for segmentation of the boundaries of corneal layers in optical coherence tomography images using Gaussian mixture model. J Med Signals Sens 2014;4:171-80.
12. Rabbani H, Kafieh R, Kazemian Jahromi M, Jorjandi S, Mehri Dehnavi A, Hajizadeh F, *et al.* Obtaining thickness maps of corneal layers using the optimal algorithm for intracorneal layer segmentation. Int J Biomed Imaging 2016;2016:1420230.
13. Golabbakhsh M, Rajaei A, Derakhshan M, Sadri S, Taheri M, Adibi P. Automated acoustic analysis in detection of spontaneous swallows in Parkinson's disease. Dysphagia 2014;29:572-7.
14. Rezaei F, Omrani MR, Abnavi F, Mojiri F, Golabbakhsh M, Barati S, *et al.* Computerized analysis of acoustic characteristics of patients with internal nasal valve collapse before and after functional rhinoplasty. J Med Signals Sens 2015;5:210-9.
15. Boersma P, Weenink D. Praat: Doing Phonetics by Computer [Computer Program]. Version 5.0.23; 2015. Available from: http://www.praat.org/. [Last accessed on 2015 Apr 15].