Sign Language Recognition Research

Study - 1

Sign Language Facts

- There are different sign languages in different countries.
- American Sign Language exist over 200 hundred years.
- Sign languages have their own **grammar** and **syntax**.
- One sign in sign language can have **multiple meanings**.
- Facial expressions define whether or not you’re making a statement or asking a question.
- Apes can learn sign language and are able to communicate with humans.
- Learning Sign Language Can Help You Improve Your Vision

http://www.deserthandtherapy.com/1058-2/

Sign Language Facts - Continues

- Brain damage affects sign language in the same way it affects spoken language.[1]
- Sign language is a visual language.[1]
- ASL uses hand shape, position, and movement; body movements; gestures; facial expressions; and other visual cues to form its words.[2]
- In the ASL, the alphabets can be demonstrated using one hand. However, in German and British Sign Languages, two hands are used.[2]

1) http://mentalfloss.com/article/13107/7-things-you-should-know-about-sign-language
2) http://www.languagesunlimited.com/blog/10-facts-sign-languages/

Sign Language Facts - Continues

- Signing is used primarily by the **deaf**, it is also used by others, such as people who can hear but **cannot physically speak**, or have trouble with **spoken language** due to some other disability.
- There are more than 137 sign languages.
- In linguistic terms, sign languages are as **rich** and **complex** as any spoken language, despite the common **misconception** that they are not "**real languages**".

Wikipedia - Sign Language

Sign Language Facts - Continues

- Sign languages, like spoken languages, organize elementary, meaningless units called phonemes into meaningful semantic units.
- Sign languages, like all natural languages, are developed by the people who use them, in this case, deaf people, **who may have little or no knowledge of any spoken language**. Sign languages are **NOT** somehow dependent on spoken languages.
- On the whole, though, sign languages are independent of spoken languages and follow their own paths of development.(BSL vs ASL) ASL resembles more with Japanese Sign Language.

Wikipedia - Sign Language

Sign Language Facts - Continues

- Sign language is visual and, hence, can use simultaneous expression.
- **Postures** or **movements** of the body, **head**, **eyebrows**, **eyes**, **cheeks**, and **mouth** are used in various combinations to show several categories of information, including lexical distinction, grammatical structure, adjectival or adverbial content, and discourse functions.
- Sign languages do not have a traditional or formal **written form**. Many deaf people do not see a need to write their own language. There is no formal acceptance of any of these **writing systems for any sign language**, or even any consensus on the matter. However there are several ways to represent sign languages in written form have been developed.

Wikipedia - Sign Language

Sign Language Facts - Continues

- Popular for hearing parents to teach signs (from ASL or some other sign language) to young hearing children. **Babies can usually produce signs before they can speak.** This reduces the confusion between parents when trying to figure out what their child wants.
- Turkish Sign language **has different grammatical structure** than Turkish language.
- 7 Haziran Türk İşaret Dili Bayramı, Türkiye İşitme Engelliler Derneği, Türkiye İşitme Engelliler Spor Federasyonu

Wikipedia - Sign Language

Sign Language Facts - Continues

- Since it is a natural language, sign language is closely linked to the **culture of the deaf**, which it originates from. Thus, knowledge about **the culture is necessary to fully understand sign language**.
- Sign language communication is **multimodal**, it involves not only hand gestures (i.e. manual signing) but also non-manual signals.
- A continuous sign language recognition system is defined as a **translation of a gesture "phrase" stream to opposite meaningful speech or text**. It is more interesting than isolated because true human signs are continuous and any isolation arising will affect communication flow

http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6222666

Research : Sign Language Recognition

Sign language is a visual language and consists of 3 major components:

- **finger-spelling**: used to spell words letter by letter
- **word level sign vocabulary**: used for the majority of communication
- **non-manual features**: facial expressions and tongue, mouth and body position

- www-i6.informatik.rwth-aachen.de/~dreuw/download/021.avi

http://www-i6.informatik.rwth-aachen.de/~dreuw/database.php#databases

Research : Sign Language Recognition

- ***Isolated Word Recognition***: recognition of one word
- ***Continious Sentence Recognition*** : recognizing sequences of gestures
- Sing language is a language just like English. It has words, phrases, sentences. So what is the **basic component** in Sign language?
  - Posture, Position, Trajectory
  - How to handle these components to understand the language?
  - Basic gestures to sentences!

Microsoft Research, 2013 - https://www.youtube.com/watch?v=fy-_k7k18Io

Research : Sign Language Recognition

- Sign Languages are expressed using postures and movements of fingers, hand, arms, fists and body, including face expressions as well. All these dynamics generate a complex problem to analyze in the computer science context, then major of systems for sign language recognition have serious limitations.
- Sign language recognition can be classified into two;
    - **First one (G1) uses electronics and devices** (like position and movement sensors, accelero-meters) to capture accurate data of fingers, hands and/or arms
    - second group (G2) uses **computer vision systems**
- G1 systems have some interesting advantages, as long as they don't use *digital cameras*, they don't depend of *illumination conditions*, and they neither need to express sign oriented to some particular spot. Besides G1 systems provide the most accurate features of position, orientation, movement and velocity of signs. **Nevertheless this kind of system has a serious disadvantage due to the permanent physical contact with the sensors**

http://file.scirp.org/pdf/ENG_2016102813314756.pdf

Research : Sign Language Recognition

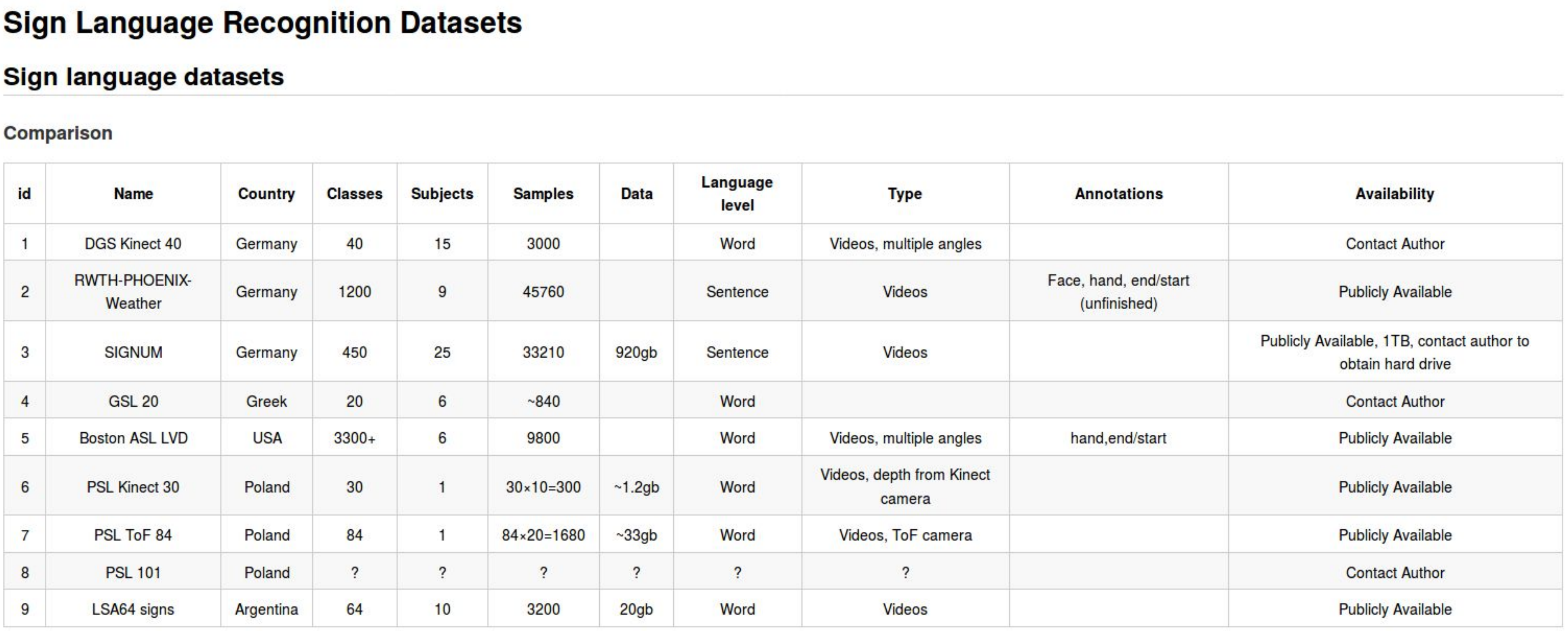
- G2 systems **allow a more natural interaction** since they don’t require signers to be connected physically to the system, but this benefit causes a considerable **loss in data accuracy**. Probably the most complex task in this kind of systems is **segmentation** (segmenting each hand from the other, hands from face, or hands from background). For their particularities, G2 systems have important limitations; **some of them have special background solid color, and some other signers use special clothes, gloves or special color markers in order to locate and segment hands.** It’s harder in these systems to calculate accurate data of position or movement of fingers, hands or some other parts of the body needed to recognize some sign, because most of systems try to solve the problem using Digital Image Processing (DIP) techniques[10]. Some of these systems use a depth sensor to improve hand shape segmentation from background; they use kinect or some other devices to capture depth information

http://file.scirp.org/pdf/ENG_2016102813314756.pdf

Research : Sign Language Recognition

- The national alphabet is part of sign language system, where each letter is represented by a **sign or gesture**, usually static, but in many languages, including Latvian, several alphabet letters are shown in motion

Review of Data Preprocessing Methods for Sign Language Recognition Systems based on Artificial Neural Networks by Zorings, Grabuts



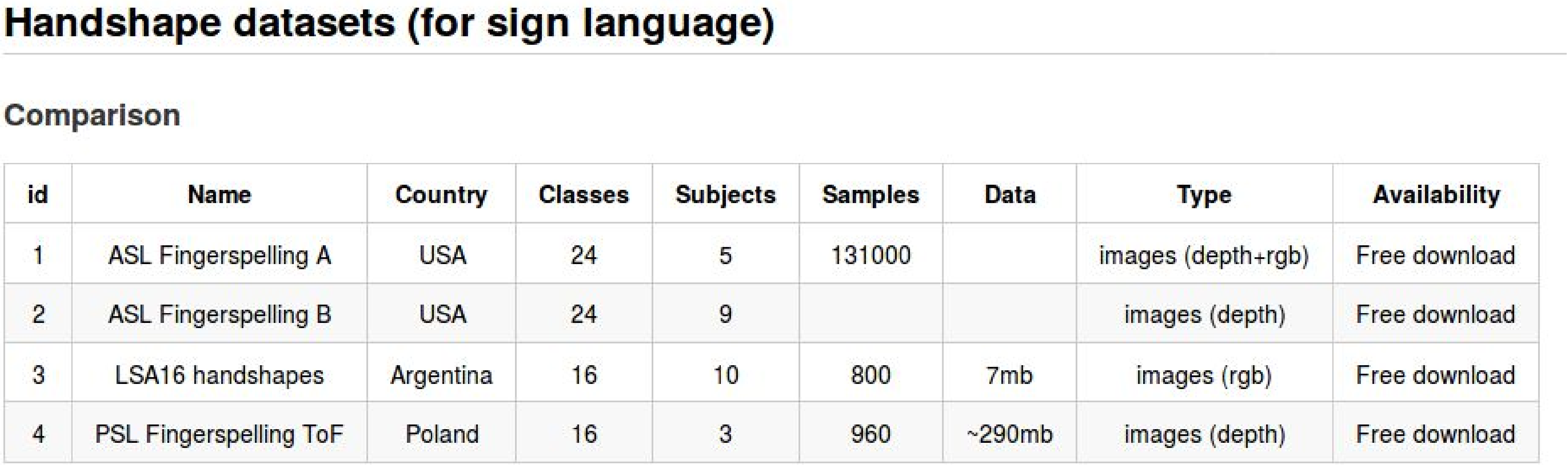
Sign Language Recognition Datasets

Sign language datasets

Comparison

id	Name	Country	Classes	Subjects	Samples	Data	Language level	Type	Annotations	Availability
1	DGS Kinect 40	Germany	40	15	3000		Word	Videos, multiple angles		Contact Author
2	RWTH-PHOENIX-Weather	Germany	1200	9	45760		Sentence	Videos	Face, hand, end/start (unfinished)	Publicly Available
3	SIGNUM	Germany	450	25	33210	920gb	Sentence	Videos		Publicly Available, 1TB, contact author to obtain hard drive
4	GSL 20	Greek	20	6	~840		Word			Contact Author
5	Boston ASL LVD	USA	3300+	6	9800		Word	Videos, multiple angles	hand,end/start	Publicly Available
6	PSL Kinect 30	Poland	30	1	30×10=300	~1.2gb	Word	Videos, depth from Kinect camera		Publicly Available
7	PSL ToF 84	Poland	84	1	84×20=1680	~33gb	Word	Videos, ToF camera		Publicly Available
8	PSL 101	Poland	?	?	?	?	?	?		Contact Author
9	LSA64 signs	Argentina	64	10	3200	20gb	Word	Videos		Publicly Available

facundoq.github.io/unlp/sign_language_datasets/index.html



Handshape datasets (for sign language)

Comparison

id	Name	Country	Classes	Subjects	Samples	Data	Type	Availability
1	ASL Fingerspelling A	USA	24	5	131000		images (depth+rgb)	Free download
2	ASL Fingerspelling B	USA	24	9			images (depth)	Free download
3	LSA16 handshapes	Argentina	16	10	800	7mb	images (rgb)	Free download
4	PSL Fingerspelling ToF	Poland	16	3	960	~290mb	images (depth)	Free download

facundoq.github.io/unlp/sign_language_datasets/index.html

Benchmark Datasets

1. **download - RWTH German Fingerspelling Database**: **German sign language,** fingerspelling, 1400 utterances, 35 dynamic gestures, 20 speakers
2. on request - **RWTH-PHOENIX** Weather Forecast: **German sign language database**, 95 German weather forecast records, 1353 sentences, 1225 signs, fully annotated, 11 speakers
3. **download - RWTH-BOSTON-50**: **American sign language database**, 483 utterances, 50 isolated signs, 83 pronunciations, 3 speakers
4. **download - RWTH-BOSTON-104**: **American sign language database**, 201 sentences, 104 signs, continuous sign language, 3 speakers, with annotated hand and head groundtruth positions for about 15k frames to evaluate tracking algorithms
5. on request - **RWTH-BOSTON-400**: **American sign language database**, 843 sentences, about 400 signs, continuous sign language, 5 speakers
6. **download - RWTH-BOSTON-Hands**: hand tracking database, 1000 frames with annotated hand positions to evaluate hand tracking algorithms - included in the RWTH-BOSTON-104 database
7. on request - **ATIS Corpus**: **Irish sign language database**, 680 sentences, about 400 signs, continuous sign language, several speakers, with annotated hand and head positions for about 5.5k frames to evaluate hand tracking algorithms
8. external - **Corpus NGT**: An online corpus of video data from **Sign Language of the Netherlands** with annotations
9. external - **BSL Corpus Project**

http://www-i6.informatik.rwth-aachen.de/~dreuw/database.php#databases

Learning Sign languages

- Turkish Sign Language
    - Sohbet ederken yüzyüze durulmalidir, konusulmalidir(dudak okunmaktadır). Jest ve mimikler, işaret dilinin temeli harf işaretleridir.
    - https://www.youtube.com/watch?v=GFHkZgmsdtE
    - https://www.youtube.com/watch?v=g_TFVtYk67s
    - https://www.youtube.com/watch?v=8M13HalSr_8&list=PLfFz63YLe29qKdNjfB5x1o_PtmbD-dHvU&index=1
- American Sign Language
    - https://www.youtube.com/watch?v=Niyz8wHXZX4
    - https://www.youtube.com/watch?v=RhQvlq-mZtA
- British Sign Language
    - https://www.youtube.com/watch?v=-2O_ymoCIR0
    - https://www.youtube.com/watch?v=dPRHENBO5ag

TRT Recordings for Turkish

- http://engelsiztrt.tv/
  - ekrana gelen dizileri, bundan böyle görme ve işitme engelliler için de düzenleyecek.
  - yayınlanacak diziler, görme engelliler için sesli olarak betimlenecek. İşitme engelliler ise, dizileri altyazılı ya da işaret dili tercümanı yardımıyla izleme olanağına sahip olacak.
- https://www.youtube.com/watch?v=312KEPlNi3s&list=PLTyi7cAWntXKR4pBElgxDiVfStwyljVN8
- http://www.trthaber.com/haberizle/isitme-engelliler/trt-haber-1400-isitme-engelliler-bulteni-2.html
- http://www.diyanet.tv/sessiz/
  - http://www.diyanet.tv/sessiz/video/sessiz-75-bolum--sevilmeye-layik-olanlar
  - Aciklayici

Videos

- Deep Learning :
    - Extending conv. Filters in time = you are not only sliding filters in space, but also in time.
        - 3D Conv. NN.[2010]
        - Sequential D.L. for Human action recognition[2011]
        - Large Scale video classification with CNNs[2014] - slow fusion, late fusion, early fusion, single frame
    - Use motion data where the small motion actually really matters a lot!! [classify swimming and tennis, properties blue vs green, motion images adds more parameters!!]
    - Two Stream CNN for action recognition in videos[Zisserman, 2014]
        - Not using 3 dimensional CNN - 2 convnets - raw images and optical flow images - fuse information at the end

Videos cont.

- We use optical flow because we don’t have many data for videos, if we have we can learn optical flow like features from raw images.
- 3D convnets so far used location motion cues to get extra accuracy(half a second or so) -
  - What if temporal dependency of interest much much larger? Several seconds
    - Speed up video?
    - Model local motion with 3d conv
    - Model global motion with LSTM - predicting classes for every single frame -- CNN + LSTM
    - Beyond short snippets DN for video classification [Ng, 2015]

Summary on video action recognition

1. Two types of architectural Patterns:
   a. Model temporal motion locally(3D Conv)
   b. Model temporal motion globally(LSTM) -- fusions of both approaches at the end
2. Clearer way get rid of RNN - make every single conv layer a RNN
   a. Delving deep into CNN for learning video representations[2016]
3. Do you think you really need a spatio-temporal video?
   a. If you really need; local vs global motion
4. Use optical flow in a second stream which can work better sometimes
5. Using different cnn & lstm methods and other methods
6. Attention models on videos(attention on frames rather than on spatial on images) - soft attention vs hard attention
7. Pretrain model on gesture dataset(20bn Gesture dataset)

Video Description

- **Describing Videos by Exploiting Temporal Structure[2015] - CODE AVAILABLE - THEANO**
  - images are static, working with videos requires modeling their dynamic temporal structure and then properly integrating that information into a natural language description. In this context, we propose an approach that successfully takes into account both the local and global temporal structure of videos to produce descriptions. First, our approach incorporates a spatial temporal 3-D convolutional neural network (3-D CNN) representation of the short temporal dynamics. The 3-D CNN representation is trained on video action recognition tasks, so as to produce a representation that is tuned to human motion and behavior. Second we propose a temporal attention mechanism that allows to go beyond local temporal modeling and learns to automatically select the most relevant temporal segments given the text-generating RNN.
- **Translating Videos to Natural Language Using Deep Recurrent Neural Networks[2015]**
  - we propose to translate videos directly to sentences using a unified deep neural network with both convolutional and recurrent structure.
- **Next Step → variational autoencoders, generative adversarial networks??**