A Real-time Gesture Recognition System for Isolated Swedish Sign Language Signs

Kalin Stefanov
KTH Royal Institute of Technology
TMH Speech, Music and Hearing
Stockholm, Sweden
kalins@kth.se

Jonas Beskow
KTH Royal Institute of Technology
TMH Speech, Music and Hearing
Stockholm, Sweden
beskow@kth.se

Abstract

This paper describes a method for automatic recognition of isolated Swedish Sign Language signs for the purpose of educational signing-based games. Two datasets consisting of 51 signs have been recorded from a total of 7 (experienced) and 10 (inexperienced) adult signers. The signers performed all of the signs 5 times and were captured with a RGB-D (Kinect) sensor, via a purpose-built recording application. A recognizer based on manual components of sign language is presented and tested on the collected datasets. Signer-dependent recognition rate is 95.3% for the most consistent signer. Signer-independent recognition rate is on average 57.9% for the experienced signers and 68.9% for the inexperienced.

1 Introduction

Sign language and different forms of sign-based communication is important to large groups in society. In addition to members of the deaf community, that often have sign language as their first language, there is a large group of people who use verbal communication but rely on signing as a complement. A child born with hearing impairment or some form of communication disability such as developmental disorder, language disorder, cerebral palsy or autism, frequently have the need for this type of communication known as *key word signing*. Key word signing systems borrow individual signs from sign languages to support and enforce the verbal communication. As such, these communication support schemes do away with the grammatical constructs in sign languages and keep only parts of the vocabulary.

While many deaf children have sign language as their first language and are able to acquire it in a natural way from the environment, children that need signs for other reasons do not have the same rights and opportunities to be introduced to signs and signing. We aim at creating a learning environment where children can learn signs in a game-like setting. An on-screen avatar presents the signs and gives the child certain tasks to accomplish, and in doing so the child gets to practice the signs. The system is thus required to interpret the signs produced by the child and distinguish them from other signs, and indicate whether or not it is the right one and if it was properly carried out.

2 Related Work

Sign languages are as complex as spoken languages. There are thousands of signs in each sign language differing from each other by minor changes in the shape of the hands, motion profile, and position. Signing consists of either manual components that are gestures involving the hands, where hand shape and motion convey the meaning or finger spelling, used to spell out words. Non-manual components, like facial expressions and body posture can also provide information during signing. Sign language recognition (SLR) inherits some of the difficulties of speech recognition. Co-articulation between signs, meaning that a sign will be modified by the signs on either side of it and large differences between signers - signer-specific styles (pronunciation in speech), both contribute to increased variation in the signing.

This paper presents a gesture recognition method that attempts to model and recognize manual components of sign language. Therefore, the work cited in this section is restricted to the specific case of tracking-based manual components extraction and modeling, and isolated word recognition using word-level classifier, where hand shape is not explicitly modeled. A comprehensive review of the research on SLR and the main challenges is provided in (Cooper et al., 2011). Manual components of sign language are in general hand shape/orientation and movement trajectories which are similar to gestures. A comprehensive survey on gesture recognition (GR) was performed in (Mitra and Acharya, 2007) and in (Rautaray and Agrawal, 2015).

Data collection is an important step in building a SLR system. Early systems used data gloves and accelerometers to capture the hands' position, orientation and velocity. These were measured by using sensors such as Polhemus tracker (Waldron and Kim, 1995) and DataGlove (Kadous, 1996), (Vogler and Metaxas, 1997). These techniques were capable of producing very accurate measurements with the cost of being intrusive and expensive. These are the main reasons for vision-based systems to become more popular. Vision-based systems can employ one or more cameras or other non-intrusive sensor (e.g. monocular (Zieren and Kraiss, 2004), stereo (Hong et al., 2007), orthogonal (Starner and Pentland, 1995), depth-sensor, such as the Kinect (Zafrulla et al., 2011)). In (Segen and Kumar, 1999) the researchers used a camera and light source to compute depth, and (Feris et al., 2004) used light sources and multi-view geometry to construct a depth image. In (Starner et al., 1998) the authors used a front view camera paired with head mounted camera. Depth can also be inferred using stereo cameras (Munoz-Salinas et al., 2008), or by using side/vertical mounted cameras as with (Vogler and Metaxas, 1998) or (ASL, 2006). There are several projects which are creating sign language datasets - in Germany, the DGS-Korpus dictionary project (DGS, 2010), in the UK, the BSL Corpus Project (BSL, 2010) and in Sweden the SSL Corpus Project (SSL, 2009). Finally, Dicta-Sign (DICTA, 2012) and SIGNSPEAK (SIGNSPEAK, 2012) are European Community's projects aiming at recognition, generation and modeling of sign language.

Hand tracking is another important part of a SLR system. Tracking the hands in sign language conversation is a difficult task since the hands move very fast and are often subject to motion blur. Furthermore, hands are highly deformable and they occlude each other and the face, making skin color based approaches complex. In early work, the hand segmentation task was simplified by colored gloves. Usually these gloves were single colored (Kadir et al., 2004). More natural and realistic approach is without gloves, where the most common detection approach uses a skin color model (Imagawa et al., 1998) and (Awad et al., 2006). Often this task is simplified by restricting the background to a specific color (Huang and Huang, 1998) or keeping it static (Starner and Pentland, 1995). In (Zieren and Kraiss, 2005) the authors explicitly modeled the background. Depth can be used to simplify the problem as in (Hong et al., 2007) and (Grzeszcuk et al., 2000). In (Fujimura and Liu, 2006) and (Hadfield and Bowden, 2012) the hands were segmented under the assumption that they are the closest objects to the camera. Recent work on multi-person pose estimation (Cao et al., 2017) illustrates a system which is capable of tracking the hands in dynamic and cluttered environments.

Early work on SLR applied Artificial Neural Networks (ANN) for modeling. The idea of one of the first papers on SLR (Murakami and Taguchi, 1991) was to train an ANN given the features from a DataGlove and recognize isolated signs. In (Kim et al., 1996) the researchers used DataGloves and Fuzzy Min Max ANN to recognize 25 isolated gestures. The work in (Waldron and Kim, 1995) presented an isolated SLR system using ANN and (Huang and Huang, 1998) presented an isolated SLR system using a Hopfield Neural Network. (Yang et al., 2002) used motion trajectories within a Time Delay Neural Network (TDNN) to recognize American Sign Language. Hidden Markov Models (HMM) (Rabiner, 1989) and (Yamato et al., 1992) are a modeling technique well suited to the problem of SLR. The work by (Starner et al., 1998) demonstrated that HMM are a strong technique for recognizing sign language and (Grobel and Assan, 1997) presented a HMM based isolated sign recognition system. (Vogler and Metaxas, 1997) showed that word-level HMM are SLR suitable. In their following work, (Vogler and Metaxas, 1999) they demonstrated that Parallel HMM (PaHMM) are superior to regular HMM, Factorial HMM and Coupled HMM for recognition of sign language.

The above mentioned systems represent important steps towards general sign language recognition.

However for the specific requirements of our application (Swedish Sign Language, real-time, modular, i.e. easy to integrate into sign-based games and extensible) none of the above systems is applicable. In the following we describe our own implementation which meets these requirements.

3 Method

Dynamic gesture recognition problem involves the use of techniques such as time-compressing templates, dynamic time warping, HMM, and TDNN. A time-domain process demonstrates a Markov property if the conditional probability density of the current event, given all present and past events, depends only on the jth most recent event. If the current event depends solely on the most recent past event, then the process is termed a first order Markov process. This assumption is reasonable to make, when considering the positions of the hands of a person through time.

An HMM is a double stochastic process governed by:

- an underlying Markov chain with a finite number of states,
- a set of random functions, each associated with one state.

In discrete time instants, the process is in one of the states and generates an observation symbol according to random function corresponding to that state. Each transition between the states has a pair of probabilities, defined as follows:

- transition probability, which provides the probability for undergoing a transition,
- output probability, which defines the conditional probability of emitting an output symbol from a finite alphabet when the process is in a certain state.

HMM have been found to efficiently model spatio-temporal information in a natural way. The model is termed "hidden" because all that can be seen is a sequence of observations. An HMM is expressed as $\lambda = (A, B, \Pi)$, where A is state transition probability, B is observation symbol probability and Π is initial state probability. For a classification problem, the goal is to classify the unknown class of an observation sequence O into one of C classes. If we denote the C models by λ_c, $1 \leq c \leq C$, then an observation sequence is classified to class c^*, where $c^* = \text{argmax}_{c \in C} P(O|\lambda_c)$. The generalized topology of an HMM is a fully connected structure, known as an *ergodic* model, where any state can be reached from any other state. When employed in dynamic gesture recognition, the state index transits only from left to right with time, as depicted in Figure 1. Here the state transition probabilities $a_{ij} = 0$ if $j < i$, and $\sum_{j=1}^{N} a_{ij} = 1$.

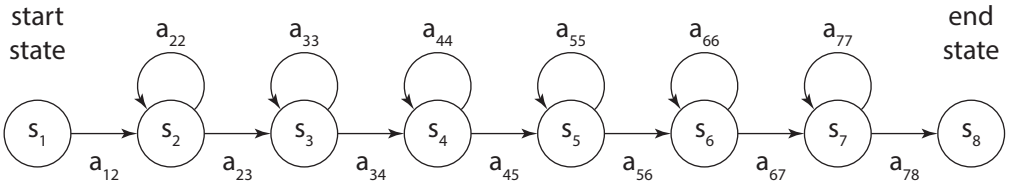

Figure 1: 8-state left-to-right HMM for gesture recognition.

The first step in the method is to train models for all 51 signs. A data recording tool was built to capture participants performing different signs in front of a Kinect sensor. The application and the collected data are described in Section 4. The feature extraction procedure (hands position) depends solely on the skeleton tracking algorithm implemented on the Kinect sensor. The results reported in this work are based on implementation that relies only on the spatial information provided from the Kinect skeleton tracking. We are working on combining skin color-based tracker and the skeleton tracker to achieve better tracking accuracy, as it is obvious that the skeleton tracker commits many errors for certain types of spatial configurations (e.g. the hands are close to each other). Furthermore, as mentioned previously, we are not modeling the shape of the hands explicitly. Let's consider the extraction of the right wrist position

feature. Visualization of the skeleton and the joints used during feature extraction is shown in Figure 2. The 3-dimensional position of the RW is calculated in user-centered coordinate space (the origin of the space coincides with the location of the signer's hip). Then the distance between the shoulders (RS and LS) is used to normalize the feature in **X**. The distance between the neck joint (C0) and the spine joint (C1) is used to normalize the feature in the **Y** dimension. The **Z** dimension of the feature is not normalized. This procedure is done for all six joints under consideration - RH, LH, RW, LW, RE, and LE. Additionally, the first derivative of each feature is calculated in order to compensate for the assumption of observation independence in the state to output probability models of the HMM. In this way a 36-dimensional feature vector is constructed and used as an observation symbol at each time step. The Baum-Welch algorithm (Rabiner, 1989) and (Rabiner and Juang, 1993) is used for training an 8-state sign-level HMM and finding the best model parameters for each sign (in the current implementation we model all signs with the same 8-state HMM topology, which might not be optimal and will be investigated further).

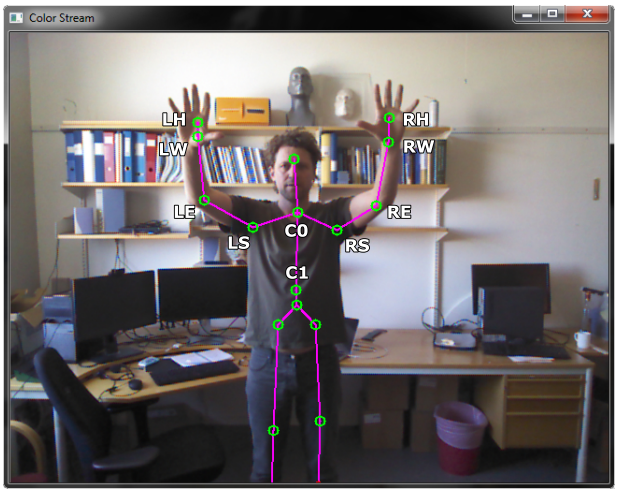

Figure 2: Visualization of the skeleton and the joints used for features extraction.

In this work the observation symbols of the HMM are the extracted feature vectors at each time step. Each class is represented by the model trained for each sign. If we get an unknown observation sequence, we have to calculate $P(O|\lambda_c)$ for all c and classify the sequence into the class which yields the maximum probability. The Viterbi algorithm (Rabiner, 1989) and (Rabiner and Juang, 1993) is applied for finding the most likely sequence of hidden states that results in the observed sequence.

4 Dataset

The size of the recorded vocabulary is 51 signs from the Swedish Sign Language (SSL). The vocabulary is composed of four subsets - objects, colors, animals, and attributes. Table 1 summarizes the vocabulary used in the current implementation.

Objects (17)	**Colors (10)**	**Animals (13)**	**Attributes (11)**
car, tractor, boat, book, ball, doll, computer, flashlight, guitar, watch, necklace, hat, key, patch, scissors, frame, drums	blue, brown, green, yellow, purple, orange, red, pink, black, white	moose, monkey, bear, dolphin, elephant, horse, camel, rabbit, cat, crocodile, lion, mouse, tiger	angry, happy, sad, small, dotted, striped, checkered, narrow, large, thick, tired

Table 1: Vocabulary.

The participants are recorded performing all 51 signs in one session. In total 5 sessions are recorded, resulting in 5 instances of each sign per participant. The recording environment is not explicitly controlled, the only requirement is that the upper body of the participant falls in the field of view of the

Kinect. Some signs involve movements that are difficult for accurate tracking with the skeleton tracking algorithm. Although this introduces errors in the dataset, we kept these instances assuming that similar errors will be committed by the algorithm during real-time usage.

4.1 Data Collection Tool

The system uses the Kinect sensor as input device. For the purpose of recording we created a data collection tool that prompts the participant to perform a certain sign and captures the color, depth and skeleton streams produced by the Kinect. The process of recording a sign follows a certain path. First the meaning of the sign is displayed as text and a video that demonstrates the sign is played, with an option of replaying the video. Once the participant is comfortable enough, the recording of the sign is started. All participants were instructed to start at rest position (both hands relaxed around the body), then, perform the sign and go back to rest position. If the sign is performed correctly the recording moves to the next one, otherwise, the participant can perform the sign until he/she is satisfied.

4.2 Experienced Signers

This part of the dataset is composed of 7 participants (5 sessions each) performing the set of 51 signs. The participants are experts in SSL (sign language researchers/teachers), where six of them are deaf. They were asked to perform the signs in the way they would normally sign. We collected total of 1785 sign instances and used all data in the experiments.

4.3 Inexperienced Signers

This part of the dataset is composed of 10 participants (5 sessions each) performing the set of 51 signs. The participants had no prior experience in SSL. They were instructed on how to perform the signs in two ways - by a video clip played back in the data recording application, and by live demonstration by the person operating the recording session. We collected total of 2550 sign instances, and used all data in the experiments.

5 Experiments

The experiments are done on the dataset described in Section 4. The results are based only on spatial features extracted from the collected skeleton data, Section 3. The conducted experiments are divided into two groups - signer-dependent and signer-independent.

5.1 Signer-dependent

In the signer-dependent experiment we test the recognition rate of the HMM only for separate signers. As described previously, we collected 5 instances of each sign for 7 experienced signers and 5 instances of each sign for 10 inexperienced signers. Table 2 summarizes the results for the experienced signers and Table 3 summarizes the results for the inexperienced signers. All results are based on leave-one-out cross-validation procedure, where the models were trained on 4 instances of each sign for each signer and tested on the 5th instance of the sign performed by that signer.

Participant	1-BEST (%)	2-BEST (%)	3-BEST (%)	4-BEST (%)	5-BEST (%)
1	92	96	97.6	99.2	99.2
2	85.5	91.8	93.3	94.9	96.5
3	84.4	92	95.6	96	96.4
4	**95.3**	97.6	98.8	99.6	99.6
5	88	93.2	95.2	96	96.4
6	87.8	94.1	96.5	98.4	98.4
7	**80**	87.5	92.9	95.3	96.5
μ	87.6	93.2	95.7	97.1	97.6
σ	5	3.3	2.2	1.9	1.4

Table 2: Signer-dependent results (7 experienced signers).

Participant	1-BEST (%)	2-BEST (%)	3-BEST (%)	4-BEST (%)	5-BEST (%)
1	84.3	92.6	96.9	96.9	97.6
2	92.2	96.1	97.3	98.8	99.6
3	**75.7**	83.9	89.4	92.5	93.7
4	94.9	99.2	99.6	100	100
5	94.5	98.4	99.2	99.2	99.2
6	93.3	96.7	97.6	98	98.8
7	89.4	96.5	97.2	98	98.4
8	**95.3**	98	99.2	100	100
9	94.1	96.5	98.4	98.4	98.8
10	89.8	96.1	97.6	98.4	98.8
μ	90.3	95.4	97.2	98.1	98.5
σ	6.2	4.4	2.9	2.3	1.8

Table 3: Signer-dependent results (10 inexperienced signers).

5.2 Signer-independent

In the signer-independent experiment we test the recognition rate of the HMM between signers. The results shown in Table 4 are based on the data from the experienced signers. We employ leave-one-out cross-validation procedure, where the HMM are trained on 30 instances of each sign (6 different signers) and tested on the 5 instances of the sign performed by the 7th signer. Figure 3 illustrates the confusion matrix for the experienced signers. The matrix is composed of the results for all 7 signers, where the shades of gray of each cell represent the recognition rate for the particular sign. The maximum point in the matrix is 1 which shows that there are signs that are fully recognizable in the signer-independent case, given the simple spatial features. On the other hand, in the figure we can also see that signs which are similar in terms of trajectories are confused (e.g. *drums* - sign #50 and *car* - sign #5).

Participant	1-BEST (%)	2-BEST (%)	3-BEST (%)	4-BEST (%)	5-BEST (%)
1	62	80.8	87.6	94.8	98
2	53.3	68.6	77.3	82.3	87.1
3	**48**	65.2	72.6	78.4	81.2
4	**65.5**	82	86.7	90.6	93.7
5	62.6	75.6	79.2	81.4	84.8
6	59.2	69.8	74.9	78.8	83.1
7	54.9	69	78	83.5	88.6
μ	57.9	73	79.5	84.5	88.1
σ	6.1	6.5	5.7	6.2	6

Table 4: Signer-independent results (7 experienced signers).

The results shown in Table 5 are based on the data from the inexperienced signers. We again employ leave-one-out cross-validation procedure, where the HMM are trained on 45 instances of each sign (9 different signers) and tested on the 5 instances of the sign performed by the 10th signer. Figure 4 illustrates the confusion matrix for the inexperienced signers. In this test none of the signs was fully recognizable between signers (maximum of 0.98). Nevertheless, the overall performance compared to the experienced signers increased with 11%.

6 Conclusion and Future Work

As expected, the performance in the signer-independent experiment is significantly lower than in the signer-dependent - 57.9% and 68.9% compared to 87.6% and 90.3% when averaged over all signers. These accuracy rates are however for the full set of 51 signs. In our current application, there is no situation where the recognizer needs to pick one sign from the full set, instead it is always the case of one out of a small number (e.g. choose one out of five animals). For this type of limited recognition tasks, accuracy will increase drastically. Furthermore, we can control the mix of signs in the games, meaning that we can make sure that signs which are confused by the recognizer never appear together - the performance of the signer-independent recognizer will not be a limiting factor in the game. This means that we can have high confidence in the recognitions and the children will be able to advance in the game only after they have learned the correct way to carry out each sign.

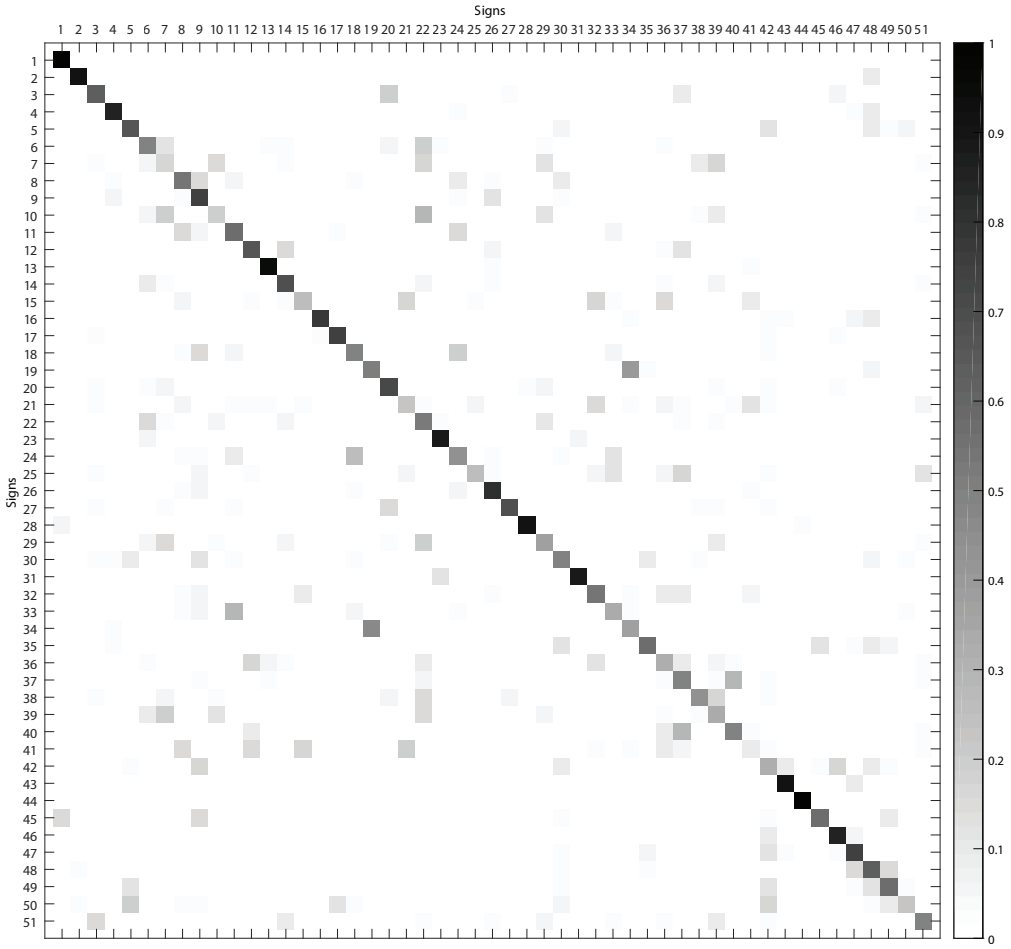

Figure 3: Confusion matrix for all experienced signers.

Participant	1-BEST (%)	2-BEST (%)	3-BEST (%)	4-BEST (%)	5-BEST (%)
1	58.8	73.7	85.9	89.4	92.5
2	72.2	84.7	89.8	91.8	94.1
3	**57.3**	73.7	81.2	88	91.6
4	69.4	80.8	87.4	91	92.5
5	75.3	91	93.7	94.9	97.3
6	**76.1**	85.9	92.5	97.6	98.4
7	73.7	83.1	90.6	91.8	94.9
8	65.5	79.6	84.3	87.4	89.8
9	**76.1**	87.4	91.8	94.5	97.3
10	64.7	79.2	88.2	90.6	94.9
μ	68.9	81.9	88.5	91.7	94.3
σ	7	5.6	3.9	3.2	2.8

Table 5: Signer-independent results (10 inexperienced signers).

An interesting observation is the fact that there is a significant difference in the recognition rate when comparing the experienced and inexperienced signers. One way to explain this is that experienced signers are more casual than their inexperienced counterpart. A similar phenomenon is also observed in speech, where non-native speakers tend to articulate more clearly than native speakers. Another major difference is the speed of signing - the experienced signers are considerably faster, which can be a challenge for the skeleton tracking algorithm.

For every sign either one hand is dominant and the other is secondary (this is not related to whether the signer is left- or right-handed) or the sign is fully symmetrical. The one hand dominance in non-symmetrical sings became obvious while studying the data from the experienced signers, where there was full consistency between signers in that respect. On the contrary, the inexperienced signers used

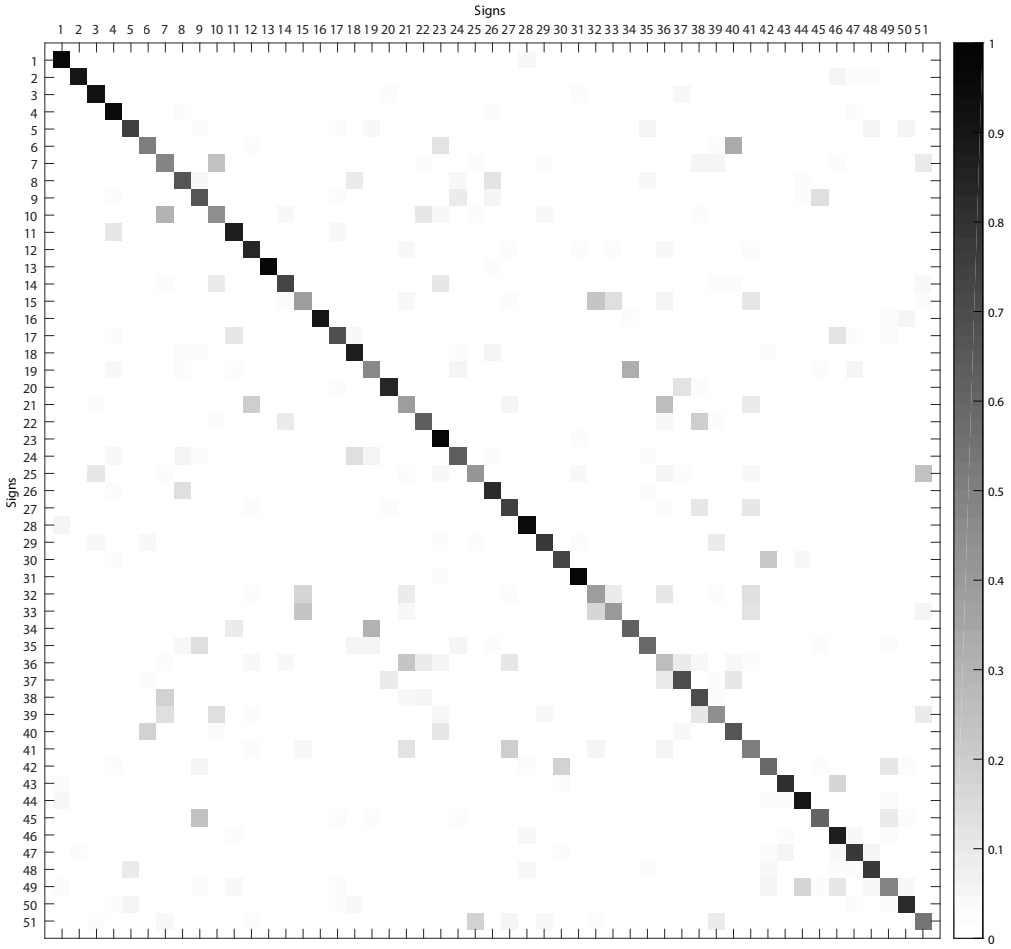

Figure 4: Confusion matrix for all inexperienced signers.

the hand that is dominant for writing, as dominant during signing. Since the target group of the system is not experienced in signing, it was decided to create models for both left- and right hand dominant versions of all signs, yielding an extended vocabulary of 102 signs, where the 51 signs in the original vocabulary are accompanied by their mirrored versions. During gameplay, once a sign is recognized to be performed *correctly* by the child but the dominant hand is swapped, we could provide feedback regarding this *mistake*.

The results obtained in these experiments show that the performance of the signer-independent recognizer is likely to be good enough for the target application. There are two main problems we plan to investigate. There is much room for improvement in the recognition accuracy. Examining the signer-dependent experiment, we can conclude that the simple spatial features used in this work are not sufficient. This is due to the fact that the skeleton tracker commits many errors in the estimates for the joints position, but also some signs are almost indistinguishable from spatial point of view (e.g. *yellow* and *orange*). We plan to extend the feature extraction procedure with a color-based hand tracker. The tracker is based on adaptive modeling of the skin color of the current user (by taking the face as reference). We expect that the hand tracker will increase the robustness of the tracking when combined with the skeleton tracker. Furthermore, we can introduce hand shape features. Further improvements are expected by introducing adaptation of the models based on a small set of signs from the target signer that could be collected during an enrollment/training phase in the game. We plan to continue recording new signs but creating a big set of isolated signs is a time consuming process.

Acknowledgments

The project Tivoli (Sign learning through games and playful interaction) is supported by the Swedish Post and Telecom Authority (PTS). We would like to thank Krister Schönström for the support during the experienced signers dataset recording.

References

ASL. 2006. http://www.bu.edu/asllrp/cslgr/.

G. Awad, J. Han, and A. Sutherland. 2006. A Unified System for Segmentation and Tracking of Face and Hands in Sign Language Recognition. In *International Conference on Pattern Recognition*, volume 1, pages 239–242.

BSL. 2010. http://www.bslcorpusproject.org/.

Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *IEEE Conference on Computer Vision and Pattern Recognition*.

H. Cooper, B. Holt, and R. Bowden, 2011. *Sign Language Recognition*, pages 539–562. Springer London.

DGS. 2010. http://www.sign-lang.uni-hamburg.de/dgs-korpus/.

DICTA. 2012. http://www.dictasign.eu.

R. Feris, M. Turk, R. Raskar, K. Tan, and G. Ohashi. 2004. Exploiting Depth Discontinuities for Vision-based Fingerspelling Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–155.

K. Fujimura and X. Liu. 2006. Sign Recognition Using Depth Image Streams. In *International Conference on Automatic Face and Gesture Recognition*, FGR'06, pages 381–386. IEEE Computer Society.

K. Grobel and M. Assan. 1997. Isolated Sign Language Recognition Using Hidden Markov Models. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 1, pages 162–167. IEEE.

R. Grzeszcuk, G. Bradski, M. H. Chu, and J. Y. Bouguet. 2000. Stereo Based Gesture Recognition Invariant to 3D Pose and Lighting. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 826–833.

S. Hadfield and R. Bowden. 2012. Generalised Pose Estimation Using Depth. In *European Conference on Trends and Topics in Computer Vision*, ECCV'10, pages 312–325. Springer-Verlag.

S. Hong, N. A. Setiawan, and C. Lee, 2007. *Real-Time Vision Based Gesture Recognition for Human-Robot Interaction*, pages 493–500. Springer Berlin Heidelberg.

C.-L. Huang and W.-Y. Huang. 1998. Sign Language Recognition Using Model-based Tracking and a 3D Hopfield Neural Network. *Machine Vision and Applications*, 10(5):292–307.

K. Imagawa, S. Lu, and S. Igi. 1998. Color-based Hands Tracking System for Sign Language Recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 462–467.

T. Kadir, R. Bowden, E. J. Ong, and A. Zisserman. 2004. Minimal Training, Large Lexicon, Unconstrained Sign Language Recognition. In *British Machine Vision Conference*.

M. W. Kadous. 1996. Machine Recognition of Auslan Signs Using PowerGloves: Towards Large-Lexicon Recognition of Sign Language. In *Workshop on the Integration of Gesture in Language and Speech*, pages 165–174.

J.-S. Kim, W. J., and Z. Bien. 1996. A Dynamic Gesture Recognition System for the Korean Sign Language (KSL). *IEEE Transactions on Systems, Man, and Cybernetics*, 26(2):354–359.

S. Mitra and T. Acharya. 2007. Gesture Recognition: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics*, 37(3):311–324.

R. Munoz-Salinas, R. Medina-Carnicer, F. J. Madrid-Cuevas, and A. Carmona-Poyato. 2008. Depth Silhouettes for Gesture Recognition. *Pattern Recognition Letters*, 29(3):319–329.

K. Murakami and H. Taguchi. 1991. Gesture Recognition Using Recurrent Neural Networks. In *Conference on Human Factors in Computing Systems*, CHI'91, pages 237–242. ACM.

L. Rabiner and B.-H. Juang. 1993. *Fundamentals of Speech Recognition*. Prentice-Hall.

L. Rabiner. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286.

S. S. Rautaray and A. Agrawal. 2015. Vision Based Hand Gesture Recognition for Human Computer Interaction: A Survey. *Artificial Intelligence Review*, 43(1):1–54.

J. Segen and S. Kumar. 1999. Shadow Gestures: 3D Hand Pose Estimation Using a Single Camera. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 485.

SIGNSPEAK. 2012. http://www.signspeak.eu.

SSL. 2009. http://www.ling.su.se/english/research/research-projects/sign-language.

T. Starner and A. Pentland. 1995. Real-Time American Sign Language Recognition from Video Using Hidden Markov Models. In *International Symposium on Computer Vision*, ISCV'95, pages 265–. IEEE Computer Society.

T. Starner, J. Weaver, and A. Pentland. 1998. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375.

C. Vogler and D. Metaxas. 1997. Adapting Hidden Markov Models for ASL Recognition by Using Three-dimensional Computer Vision Methods. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 1, pages 156–161.

C. Vogler and D. Metaxas. 1998. ASL Recognition Based on a Coupling Between HMMs and 3D Motion Analysis. In *International Conference on Computer Vision*, pages 363–369.

C. Vogler and D. Metaxas. 1999. Parallel Hidden Markov Models for American Sign Language Recognition. In *International Conference on Computer Vision*, pages 116–122.

M. B. Waldron and S. Kim. 1995. Isolated ASL Sign Recognition System for Deaf Persons. *IEEE Transactions on Rehabilitation Engineering*, 3(3):261–271.

J. Yamato, J. Ohya, and K. Ishii. 1992. Recognizing Human Action in Time-sequential Images Using Hidden Markov Model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–385.

M.-H. Yang, N. Ahuja, and M. Tabb. 2002. Extraction of 2D Motion Trajectories and Its Application to Hand Gesture Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1061–1074.

Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti. 2011. American Sign Language Recognition with the Kinect. In *International Conference on Multimodal Interfaces*, ICMI'11, pages 279–286. ACM.

J. Zieren and K.-F. Kraiss. 2004. Non-intrusive Sign Language Recognition for Human-Computer Interaction. In *Symposium on Analysis, Design and Evaluation of Human Machine Systems*, page 27.

J. Zieren and K.-F. Kraiss, 2005. *Robust Person-Independent Visual Sign Language Recognition*, pages 520–528. Springer Berlin Heidelberg.