The Automatic Recognition of Isolated Sign Language Signs Based on Gesture Components and DTW Algorithm

Katarzyna Barczewska
AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Krakow, Poland
Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering
Department of Automatics and Biomedical Engineering
e-mail: kbarczew@agh.edu.pl

Author presents sign language features that can provide the basis of the sign language automatic recognition systems. Using parameters like position, velocity, angular orientation, fingers bending and the conventional or derivative dynamic time warping algorithms classification of 95 signs from the AUSLAN database was performed. Depending on the number of parameters used in classification different accuracy values were obtained (defined as the ratio of correctly recognized gestures to all gestures from test set), with the highest value 87.7% for the case of classification based on all the features and the derivative dynamic time warping method.

Keywords and phrases: automatic gesture recognition, sign language, DTW, DDTW.

Background

With the technology development we can observe more interesting solutions to human-computer interfaces. A few years ago the primary interface that was used was a mouse and a keyboard, nowadays touch screens, webcams, sensors, Microsoft Kinect or Leap Motion provide new ways of communication with computer and gradually replace former devices giving user completely new possibilities: controlling the application by touching the screen, performing different gestures. Advantages of these devices can be used not only in computer games but also as a help in communication for deaf or impaired people. Taking into account all these aspects, author presents sign language parameters that can be recognized using these devices in automatic recognition of sign language gestures and the results of AUSLAN database gestures recognition performed using the manual parameters of sign language and methods of the time series alignment.

The natural language and way of communication of deaf and hard of hearing people is sign language. However it has different form, it enables people to express the same things as a spoken language. Known by deaf people all over the world, differs regionally, and exists in many national versions American, German, Polish and other national sign languages [1,2].

Contrary to that majority of hearing people think, it has completely different grammar, structure than any other national spoken language [2]. Grammar aspects are not well understood by hearing people, that is why it is still a big challenge for them to learn this language – as well as for deaf people to learn spoken language to write or read anything in spoken languages. Thanks to online translators it is possible for everybody to translate words, phrases from almost any foreign language to his/her mother tongue. Thanks to all of the devices mentioned in introduction we are closer to translate sign language gestures into text or speech and enable communication between deaf and hearing people. There are research projects all over the world which goal is to recognize and translate sign language into the speech, examples are: continuous gesture recognition of Taiwanian Sign Language [3] automatic recognition of isolated gestures of Danish Sign Language [4], research on syntactic aspects of American Sign Language [5], continuous recognition of American Sign Language sentences [6], translation from sign language into speech and from speech into sign language [7]. The subject of the analysis are visual signals, images that have completely different characteristics than speech (acoustic) signal. There are two means of expression in sign language: the manual and non-manual channel. Manual channel include hand gestures and non-manual consists of facial expressions, movements of head and whole upper body. The first channel is used to express lexical meaning of sign language words, while the second one, co-occuring with the gestures, provides syntactic information in the utterances, grammatical markers into sentences. Using facial expressions signer can also express emotions, attitudes, using whole body can imitate actions [5].

Important step, from automatic recognition systems' point of view, is the distinction of parameters which describe gestures. Each gesture in sign language consists of

a few components, that can be among the manual means of expression: hand shape, hand posture, hand location, and hand motion, among non-manual: head and body posture, facial expression, gaze and lip patterns [1].

In [3] authors distinguish four components of Taiwanian Sign Language (taking into account only hand gestures), which are: posture, position, orientation, and motion. Posture is defined as specific hand flexion configuration observed at some time instance. Position parameter corresponds to the hand location in relation to whole body, e.g. above the head, close to the ear, close to the heart, in the front of the face. Orientation parameter refers to both palm and fingers, there are words in Taiwan Sign Language which differ only in the index finger orientation or in the palm orientation. The last parameter is motion trajectory: movement can be linear, circular, U-like, L-like, J-like, with arm waving, wrist waving or with wrist rotation. The gesture is a sequence of postures performed in specific orientation connected by motions over a short time span. Authors distinguish 51 fundamental postures, 22 basic gesture/body relative positions and 8 different motion types [3]. In [1] authors describe characteristic areas of interest, that are important in the recognition of facial expressions, these parameters describe shape and the position of the eyes, eyebrows, and mouth (in particular the lips) as well as their spatial relation to each other.

In this article author focuses on identifying isolated sign language gestures basing on the manual components, similar to the described above: hands' position, movement velocity, angular orientation of palms, fingers bending. Algorithms used in classification are conventional and derivative dynamic time warping that have been successfully applied in recognition of simple gestures in previous research described in [8] and [9], as well as in research presented in [10] or [11].

Material and Methods

Database

In the study AUSLAN database was used. Prepared and described in [12], [13], database consist of 95 different gestures of Australian Sign Language performed by one volunteer native Auslan signer in 9 measurement sessions over a period of nine weeks. Each gesture was repeated 3 times during one session, so 27 samples of each sign were collected, giving 2565 signs in total. Data was captured using a setup of data gloves and magnetic position trackers attached to each hand. Gloves provided information about fingers bending (measured between 0 and 1, 0 – totally flat, 1 – totally bent), magnetic trackers – about angular orientation of hands (roll, pitch and yaw angles) and their position in relation to human body (x, y, z, with the origin of coordinate system below the chin). Sampling rate of complete system was 100 frames per second, average length of each gesture was about 57 frames [12]. Database was divided into two sets: training and test set. The training set contained 2/3 of all gestures (18 repeats of each sign), the test set 1/3 (9 repeats of each sign).

Algorithms

At first all gestures signals were filtered to eliminate the effect of shaky hands. In the next step segmentation was carried out to indicate time boundaries of each gesture. Segmentation based on the changes of standard deviation counted in windows containing small number of frames. Significant increment of the value of standard deviation calculated in window was considered as the boundary. Because of the fact that database was collected for one person's gestures only, there was no need to normalize amplitudes of the signals. The same gestures performed by different people very often differs in many features, e. g. in ranges of the parameter values, in the dynamic characteristics. Taking into account repeats of gestures made by one person, it can be observed that they differ too, but not so significantly. Depending on many factors like person's attitude, time of the day, tiredness the same gestures can be made with different verve, speed, force.

The algorithms used for gestures comparison and classification were Dynamic Time Warping (DTW) and its modified version, Derivative Dynamic Time Warping (DDTW). These algorithms were chosen because of specific characteristic of gestures. Both allow the assessment of the similarity of the time series. They are used to compare signal patterns that are similar but transformed in time [9], [14].

Dynamic Time Warping

DTW algorithm is used to compute the distance between two signals. Signals don't have to be the same length. The procedure is as follows: $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\}$ if and are two gesture signals (given by any parameter like position, velocity, angle signals changing in time), to align these two sequences distance matrix D with Euclidian distances between all pair of points (x_i, y_j) has to be defined:

$$D(i,j) = d(x_i, y_j) \tag{1}$$

where:

$$d(x_i, y_j) = |x_i - y_j| \tag{2}$$

for 1D signals, and:

$$d(x_i, y_j) = \sqrt{(x_{i1} - y_{j1})^2 + (x_{i2} - y_{j2})^2 + (x_{i3} - y_{j3})^2} \tag{3}$$

for 3D signals, where $x_i = (x_{i1}, x_{i2}, x_{i3})$ and $y_j = (y_{j1}, y_{j2}, y_{j3})$. Cumulative matrix P is defined:

$$P(1,1) = 0;$$
$$P(i,1) = D(i,1) + P(i-1,1);$$
$$P(1,j) = D(1,j) + P(1,j-1);$$

For $i, j > 1$
$$P(i,j) = D(i,j) + \min\{P(i-1,j), P(i,j-1), P(i-1,j-1)\}$$

At the end parameter q_{DTW} is defined, which is the result of the DTW algorithm and specifies the optimal total distance between X and Y after alignment [15], [9], [8]. Illustration of the time sequences alignment using DTW algorithm is shown on Figure 1.

Derivative Dynamic Time Warping

DDTW algorithm, described by Keogh and Pazzani [17], gave better results than DTW in simple gestures recognition research [9]. It is useful when signals that are being compared have local differences in amplitude. Algorithm looks the same like DTW described above, the difference is that instead of the signals X and Y the input to the algorithm are their derivatives, X' and Y'. The estimate of the derivative was calculated for each point of the discrete signal by Eq.(4) [17]:

$$x'_i = \frac{(x_i - x_{i-1}) + ((x_{i+1} - x_{i-1})/2)}{2} \quad (4)$$

Where x is an element of vector X with index i. The result of DDTW is parameter $q = P_{DDTW}(n,m)$.

Recognition

Analysis of signals was performed in application written for the purpose of this research in visual programming language LabVIEW. The manual components of gestures that were taken into consideration in the classification were:

- hands position – 3 components: x, y, z, for each hand;
- angular orientation of the palms – 3 components: yaw, pitch and roll for each hand;
- fingers bending – 1 component for each finger, 5 components for each hand;

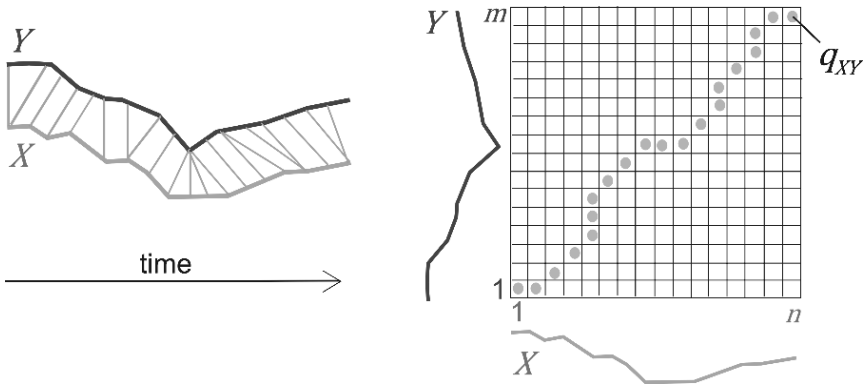

Figure 1. Illustration of time sequences alignment using DTW algorithm. Sequence X contains n elements, Y – m elements. The orange path on the matrix P is the warping path which connects pairs of points from X and Y sequences, between which distance has the smallest value. The last element of the path is equal to $q = P(n,m)$. Illustration based on [16].

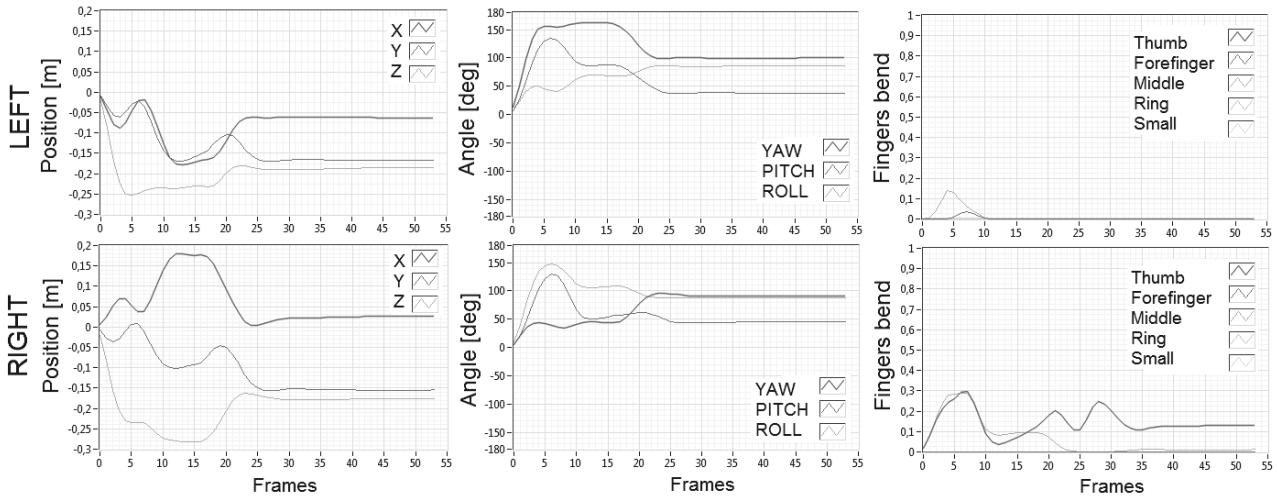

Figure 2. All components for the sign lose

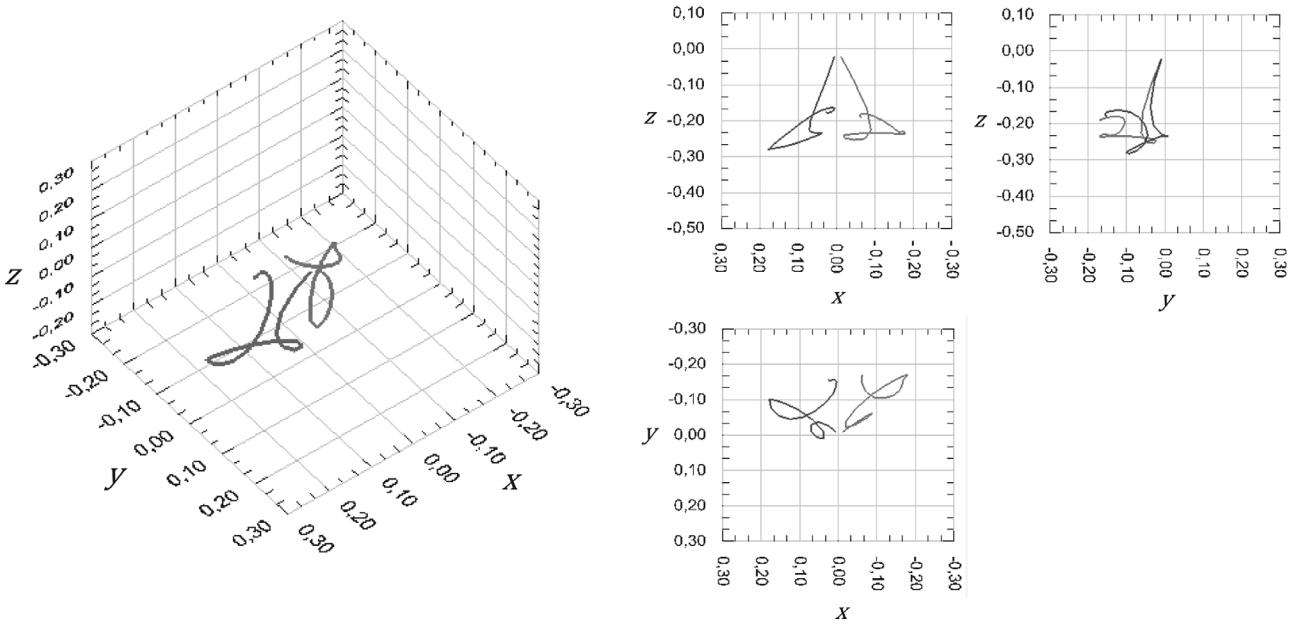

Figure 3. Hands motion trajectories in 3D space for gesture lose. Red color corresponds to the right and green to the left hand

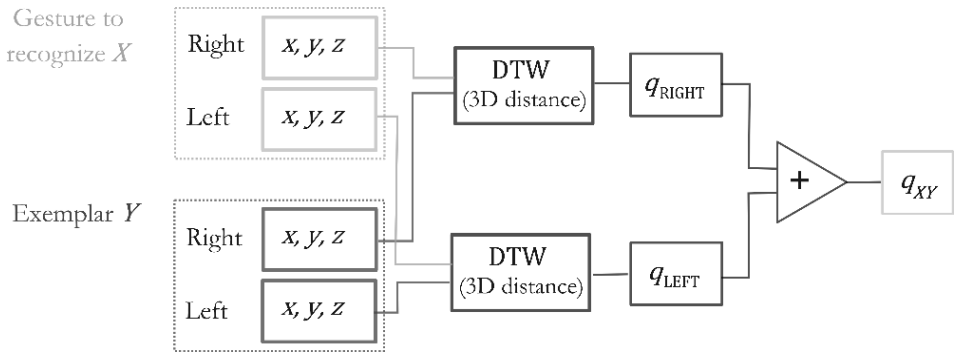

Figure 4. Calculation of DTW distance q_{xy} between hands position in 3D space: matching of the gesture X from the test set to the exemplar Y

- velocity of hands movement, calculated from position – calculated for each of 3 position components for each hand.

Thus each gesture consists of 3 + 3 + 5 + 3 = 14 signals for one hand. Classification is based on two hands, using 28 signals to recognize one sign. Example of filtered components for the sign *lose* are presented on Figure 2. Hands motion trajectories in 3-dimensional space for the same sign are shown on Figure 3. The sign *lose* was chosen to demonstrate the character of different components, because during its performance almost all of them change over the time.

Similarly to the research and recognition of simple gestures described in [8], [9], the classification covers comparison to the exemplars which were indicated in the training set. For each of 95 AUSLAN signs one sign was indicated as an exemplar. The gesture from training set became an exemplar, when it was the most similar to other repeats of the same sign in terms of one of the warping methods DTW or DDTW. In other words similarity was measured by DTW or DDTW algorithm, and 18 repeats of the same sign were compared to each other, q parameters obtained for each comparison were summed. The gesture that's total parameter q had the smallest value became an exemplar.

In the classification step, gestures from the test set were aligned to each of the exemplars using DTW or DDTW algorithms. The gesture from the test set was recognized as the sign corresponding to exemplar whose parameter q had the smallest value. Scheme of signal and exemplar alignment is presented on Figure 4.

Results and discussion

Signs classification was performed using two warping algorithms: DTW, DDTW, the test set and the sets of exemplars. Recognition was carried out for single parameters, for sets of different parameters and for all of them at once. Position, velocity and angles were treated as 3-dimensional signals and the distance in DTW algorithms was calculated by Eq. (3) in 3-dimensional space giving one value for all 3 components for each of these parameters. Signals with information about fingers bending were treated as 1D, and the distance was calculated for each finger separately by Eq. (2). To identify a sign q parameter was calculated for both hands, and in the classification was used the sum of q obtained for right and for left hand (like presented on the scheme on Figure 4) For the recognition that was based on the set of parameters, DTW distance q was calculated by summing all q's obtained for all single parameters, and then the exemplar that has the smallest total value was indicated as recognized one.

As the result, accuracy of the classifier was obtained for the test set. It was calculated as the sum of all correctly recognized gestures, n_c (the sum of the true positives for all classes), divided by the sum of all gestures from test set,

$$A = \frac{n_c}{n_{TOTAL}} \times 100\% \quad (5)$$

Accuracy values for each of the described recognition method and for different basis of classification are shown in Table 1.

Comparing values of the accuracies shown in Table 1, it can be observed that the more parameters taken into classification, the better recognition results. For the single parameter recognition, classification based only on fingers bending gave very poor results. Better recognition was achieved for angular orientation of palm, position or velocity, but still not high enough. It turned out that the recognition results become significantly increasing while adding more parameters to the basis of recognition. Thus, the highest value of the recognition accuracy was obtained when the basis of classification was the set of all parameters, for both the DTW and the DDTW algorithms. It confirms the theory about sign gestures components outlined in the introduction. For different gestures the same component can be identical, but there is at least one, like bending of special finger, or position in relation to body, which allows the identification of proper sign. Both used algorithms allow to obtain satisfactory classification results. Previously they were successfully used in the recognition of the set of 10 simple gestures [9], but as has been demonstrated they also work well with larger number of more complex gestures. As in the case of simple gestures classification, better results are obtained while using DDTW algorithm – accuracy is then 87.7%, for DTW algorithm its value is slightly lower, is 85.6%. Adding more parameters like acceleration, distance between hands or angular velocity could be helpful in the recognition process and cause increment of the recognition accuracy.

Table 1. Recognition accuracy (all values are in %). Results for DTW and DDTW algorithms used with different input parameters and their groups

Basis for classification	DTW [%]	DDTW [%]
Position	33.5	63.6
Velocity	63.6	41.9
Angles	60.0	68.8
Fingers bending	34.0	36.3
Pos + Ang	60.8	72.9
Pos + Ang + Fing	63.5	69.9
Pos + Ang + Vel	76.1	79.5
Pos + Ang + Vel + Fing	85.6	87.7

Conclusions

The algorithms DTW and DDTW were used to classify 95 signs from AUSLAN database. Database described in [12], [13] consisted of signals with information about hands position, angular orientation of palms, fingers bending. Additionally the velocity of hands movement was calculated using position parameter. The set of all gestures was divided into the training and the test set. Among the gestures from the training set, exemplars were indicated for each of 95 signs. Gesture became an exemplar when it was the most similar to other repeats of the same sign in the terms of DTW or DDTW method. In the classification stage gestures were compared with exemplars using the same warping methods. DTW and DDTW methods were chosen to classification because of the characteristic of collected gesture signals which are quite similar but transformed in time. Signals for the same gestures differs in their length, distribution of extremes and the length of particular gesture phases. The best recognition was obtained for DDTW method and the basis of classification consisted of all mentioned parameters. It turned out that algorithms that are commonly used to find the similarity between time series, worked well in the automatic recognition of sign language signs. Automatic sign language recognition based on gesture components could give better results with more components, like acceleration, angular velocity or changes of the distance between hands during sign performance, calculated from the parameters mentioned above.

An interesting topic for future research is to find more different gestures components and identify among them the smallest number of such components that enable gesture recognition with sufficient effectiveness.

References

[1] von Agris U., Knorr M., Kraiss K.-F., *The Significance of Facial Features for Automatic Sign Language Recognition*, Proceeding of: Automatic Face & Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on. Amsterdam, Netherlands.

[2] Sacks O., Zobaczyć głos. Podróż do świata ciszy. Zysk i S-ka Wydawnictwo s.j., Poznań 2011.

[3] Liang R.-H., Ouhyoung M., *A Real-time Continuous Gesture Recognition System for Sign Language*, Third IEEE International Conference on Automatic Face and Gesture Recognition, Proceedings, 1998.

[4] Lichtenauer J. F., Hendriks E. A., M. Reinders J.T., *Sign Language Recognition by Combining Statistical DTW and Independent Classification*, IEEE Transactions on Pattern Analysis and Machine Inteligence, vol. 30, no. 11, 2008.

[5] Nguyen T.D., Ranganath S., *Facial expressions in American sign language: Tracking and recognition*, Pattern Recognition 45, 2012.

[6] Gweth Y. L., Plahl C., Ney H., *Enhanced Continuous Sign Language Recognition using PCA and Neural Network Features*, Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. Providence, Rhode Island.

[7] http://www.kinectranslator.com/pl/technologia/, date of last visit: July 2013.

[8] Barczewska K., Drozd A., Folwarczny Ł., *Rozpoznawanie gestów z wykorzystaniem czujników inercyjnych o 9 stopniach swobody - Gesture recognition based on 9DOF inertial sensor*, Pomiary, Automatyka, Kontrola, vol. 59, 2013.

[9] Barczewska K., Drozd A., *Comparison of methods for hand gesture recognition based on Dynamic Time Warping algorithm*, Proceedings on FedCSiS, Kraków, September 2013.

[10] Akl A., Feng C., Valaee S., A Novel Accelerometer-Based Gesture Recognition System, Transactions on Signal Processing, IEEE, vol. 59, No. 12, December 2011.

[11] Hussain S.M.A., Harun-ur Rashid A.B.M.: User Independent Hand Gesture Recognition by Accelerated DTW, IEEE/OSA/IAPR International Conference on Informatics, Electronics & Vision, Proceedings, Bangladesh 2012.

[12] http://archive.ics.uci.edu/ml/datasets/Australian+Sign+Language+signs+%28High+Quality%29, date of the last visit: July 2013.

[13] Kadous M. W., *Temporal Classification: Extending the Classification Paradigm to Multivariate Time Series*, PhD Thesis (draft), School of Computer Science and Engineering, University of New South Wales, 2002.

[14] Helwig N. E., Hong S., Hsiao-Wecksler T., *Time-Normalization Techniques for Gait Data*, 33rd Annual Meeting of American Society of Biomechanics Materials, State College, PA, USA, 2009.

[15] Müller M.: *Information Retrieval for Music and Motion. Chapter 4: Dynamic Time Warping*. Springer Verlag 2007.

[16] Tsiporkova E., Dynamic Time Warping Algorithm for. PPT presentation available at: http://www.psb.ugent.be/cbd/papers/gentxwarper/DTWAlgorithm.ppt, date of the last visit: July 2013.

[17] Keogh, E., Pazzani, M., *Derivative Dynamic Time Warping*. In First SIAM International Conference on Data Mining (SDM' 2001), Chicago, USA.