2024 21st International Conference on Mechatronics - Mechatronika (ME) | 979-8-3503-9490-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ME61309.2024.10789742

Automation-Driven Dataset Preparation for Continuous Czech Sign Language Recognition

Jan Snajder
Institute of Solid Mechanics, Mechatronics and Biomechanics
Brno University of Technology
Brno, Czech Republic
jan.snajder@vutbr.cz

Jiri Krejsa
Institute of Solid Mechanics, Mechatronics and Biomechanics
Brno University of Technology
Brno, Czech Republic
krejsa@fme.vutbr.cz

***Abstract*— This paper presents an automation-driven solution for preparing a continuous Czech Sign Language dataset, addressing the lack of resources in this area. Manual processing of daily sign language news recordings would be extremely time-consuming, as the videos vary in quality, use different overlays, and have no captions. To streamline this process, we use the Structural Similarity Index Measure (SSIM) to compare key frames and extract relevant parts of the recording, such as weather forecast segments. Automatic speech recognition (ASR) then processes the accompanying audio and generates textual transcriptions of the spoken content. The outcome is the highly automated preparation pipeline and the dataset containing 4699 annotated videos of weather forecast news in Czech Sign Language providing a foundation for future research in sign language recognition.**

Keywords—sign language, continuous, dataset, recognition, translation

I. Introduction

Sign language is the natural means of communication used primarily in the hearing-impaired community. It can be understood as a word-based system, where each sign can be roughly translated to a spoken word or a concept, known as *gloss*. Individual signs usually join hand gestures and facial expressions (or even body movement). Similarly to its spoken counterpart, there is no universal sign language, but different regions and countries have their versions of the language.

From the point of view of its recognition, sign language can be split into two parts:

- Fingerspelling, a means to sign an individual letter, is equivalent to spelling in spoken language. It is usually utilized in cases where a word does not have a specific sign yet. It can also be helpful during sign language teaching as it is an easy way to communicate if one does not know or fails to remember a specific sign.
- Sign gestures, a means to sign a whole word or a gloss. It is a core part of any sign language used most of the time. It has its own grammar and lexicon, both entirely different than in spoken languages.

The recognition of the former can be defined as a static image data classification. Most signed letters in the Czech Sign Language alphabet are characterized by static fingers and hand position, hence the task is ideal for convolutional neural networks (CNN). The dataset for the Czech Sign Language single-hand alphabet has been introduced in [1] together with classification utilizing CNNs. The classification accuracy was later improved in [2] by leveraging the framework *MediaPipe Hands* [3]. Finally, the dataset was extended by diacritics, which are expressed by hand movement – this changes the task to image sequence classification. This extension and classification method using a *Long Short-Term Memory* (LSTM) neural network was introduced in [4].

The task of recognition of sign gestures is much more complex and consists of the classification of hands and finger movements. In this case, the number of classification classes is not fixed (for fingerspelling it was the number of letters in an alphabet), and it solely depends on the dataset and its vocabulary. There are several state-of-the-art datasets of isolated signs, the most influential being *Ankara University Turkish Sign Language* (AUTSL) [5] consisting of 38,336 sequences of 226 signs, and *Word-Level American Sign Language* (WLASL) [6] containing 68,129 sequences of 20,863 signs. Both datasets are heavily used in various benchmarks, proving that current recognition methods are capable of such a task. In our previous work, we successfully tested the *MediaPipe* framework with the latter dataset [7].

Although the recognition of isolated sign language gestures seemed to be a solved task, it does not cover the grammatical and linguistic structures of the language. Most of the recognition approaches assume one-to-one mapping, however, that does not provide a meaningful interpretation of sign language thus making it unusable in real-life application. This incapability was mentioned and addressed in [8], as the authors came up with the term *"sign language translation"*, proposing sequence-to-sequence translation. Together with their method, they introduced a large-scale continuous dataset of German Sign Language – *RWTH-PHOENIX-Weather 2014T*. It consists of recordings of weather forecasts, including 7096 segments. Each segment represents one sentence or phrase in sign language and is annotated by gloss and spoken representation. Gloss annotations allow the usage of two translators: sign to gloss and gloss to words.

This paper focuses on the automation-driven preparation of a continuous Czech Sign Language dataset as there is no such resource for Czech Sign Language. Since the outcome of [9] proved that the gloss representations of sign language sequences are unnecessary and translation to spoken language from its signed counterpart directly has similar results, the whole preparation of dataset for such a task could be automated. The outcome of this paper is a pipeline, which allows us to cut out and annotate the coveted part of the video. To demonstrate the usage, we used the method to create a continuous Czech Sign Language dataset with a focus on weather forecasts. Sequences of our dataset are several sentences long, allowing researchers to put up to test the most modern machine learning models.

The organization of this paper is as follows: in the beginning, we discuss the raw data acquisition, its properties, and challenges. Then the automation of the preparation pipeline is introduced and the flow of its components, segment

This research was funded by Faculty of Mechanical Engineering, Brno University of Technology under the projects FSI-S-23-8334: "Research and development of new methods and approaches in the field of modelling and control of mechatronic systems".

979-8-3503-9490-0/24/$31.00 ©2024 IEEE

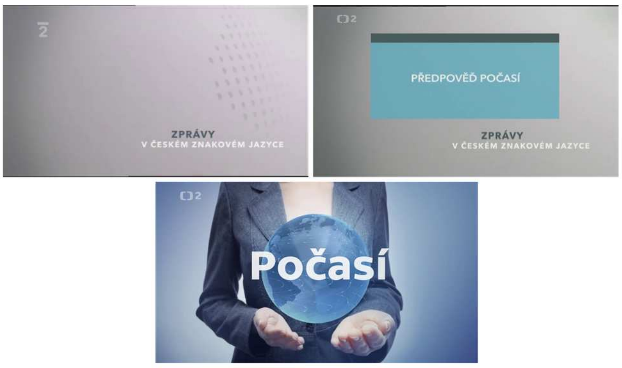

Fig. 1. Examples of different images of transition to the weather forecast segment.

extraction and annotation, are presented. Finally, we analyze the results of individual parts and the dataset itself.

II. Data Acquisition

Czech Sign Language news is part of a daily broadcast of Czech national television. Thanks to their archive, available online on *www.ceskatelevize.cz/ivysilani*, nearly 15 years of this news are publicly accessible. With the help of 3rd party software, we have gathered 4713 recordings dating from the 2nd of January 2009 till the 31st of December 2023—the news is unavailable on weekends and public holidays.

The recordings do not contain transcriptions nor captions, although for sign language translation a transcription is necessary as it has the role of a label for a given recording. To fulfill the goal of an automated approach, the use of speech recognition is indispensable. Another challenge for automation is the fact, that the overlay of the news and transitions between individual segments changed several times over the years. Different segment transitions are shown in Fig 1.

Most recordings are approximately 10 minutes long, although in earlier days news was shorter with some outliers approaching the 3 minutes mark. Each recording consists of three segments – general news, sport, and weather forecast, an average timeline of these segments is visualized in Fig. 2. Many recordings are cut incorrectly and include several minutes of commercials or even the next show. The segment of our interest was the weather forecast, however, the method presented in this paper could be used on other segments as well.

The quality of the samples varies, while in earlier days, the highest resolution available was 480p (640x480), since 2016 all recordings have been in 1080p (1920x1080). We downscaled all videos to 720p (1280x720) except those, which are already in lower quality, as that allowed us to keep relatively high quality while maintaining reasonable file sizes. Our aim was not to unify the resolution since it would significantly downgrade the quality. Further downscaling or unifying can be done by researchers to make the input suitable to their methods.

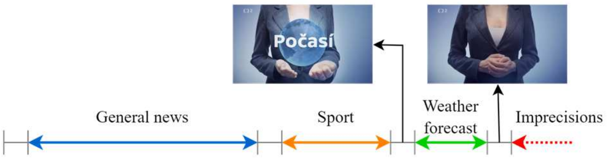

Fig. 2. Typical timeline of sign language daily news segments of downloaded recordings.

The focus of a continuous sign language dataset shall be restricted vocabulary with minimal use of unique words and names, as developing and testing various algorithms is more predictable in a minimized and controlled environment. General news does not meet any of these targets, their topics range from home politics to natural disasters all around the world and consist of many unique names and places. Sports have restricted vocabulary; however, it is still quite large and there is a frequent use of various names of people, places, or teams. That is why weather forecast has been chosen as a topic of interest, similarly to [8]. The vocabulary is very restricted, and the use of unique names is minimal. Furthermore, it is usually the shortest segment. These properties make it an ideal segment for the initial development of sign language translation.

III. Preparation Pipeline

The input data to the preparation pipeline are described in the previous chapter. The main target of the pipeline is to extract the weather forecast segment from each recording and generate a transcription of the speaker in the recording, the whole process is visualized in Fig. 3. Additionally, the pipeline is also used for the elimination of outliers. There are several either fully faulted recordings or recordings with invalid audio. To manually check thousands of videos would be extremely time-consuming. All outliers, which did not align with the pipeline are logged allowing users to manually check them later, reducing the number of videos to control.

A. Segment extraction

Since the weather forecast segment is only a fraction of the daily news recordings, it must be extracted. Segments of the news are usually separated from each other by a graphical transition. For all recording the weather forecast comes as the last segment following sports. The recording usually ends right after the weather forecast; however, it was found out that on many occasions the cut is not as precise as expected and cannot be relied on. Moreover, the preparation method that cuts the segment from both sides could be used in the future (for example for the sports segment).

The initial idea was to use optical character recognition (OCR) since there is an announcing text during the transition to the weather forecast segment. However, the fact, that there is a need to cut the recording from both sides made this

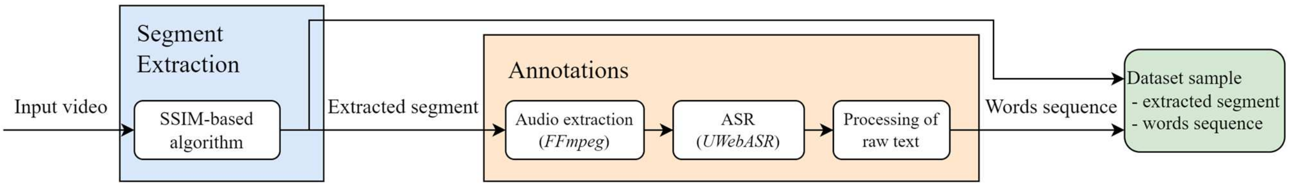

Fig. 3. The flow of the preparation pipeline.

approach unusable, because in some overlays there might not be any suitable text to look for at the end. Furthermore, OCR algorithms are often used only on certain sections of the screen, which would require further input from the user. It would be very computationally heavy to use the OCR algorithm on the entire screen and may drastically slow down the process.

Since the OCR turned out to be unsuitable for the task, an approach using image comparison to apply it to whole frames was considered. Structural similarity index measure (SSIM) [10] algorithm has been chosen and as the name suggests, it measures similarity between two images by focusing on structural information rather than pixel-by-pixel differences. It compares three aspects of an image: luminance, contrast, and structure. It is usually calculated using a sliding window approach, where these tree components are computed over local regions and combined into a single score. The score is in intervals from -1 to 1, where the latter means perfect similarity, zero indicates no similarity, and negative values suggest a structural mismatch.

The advantage of this approach is that it is possible to identify a frame without a text on it while maintaining relatively low computational demands compared to the OCR approach. Moreover, the output can be transformed into a percentage of similarity, and this information can be used in further processing and optimizations. The downside is that the approach requires a user to input images of the initial and end frames of the target segment. These images must be the same resolution as the recordings.

As mentioned in the previous paragraph, the output of the SSIM algorithm can be transformed to a percentage of similarity and used to optimize the extraction of the segment. We demonstrate such an optimization for the extraction of weather forecasts. Instead of checking every frame for similarity with our target images, we leverage the fact that every transition in the news is several seconds long and that the weather forecast is the last segment of the news.

The logic of segment extraction is shown in Fig 4. The offset is calculated based on the following equation:

$$o = TotalFrames - (m * 60 * FPS) \qquad (1)$$

The o describes offset from start of the recording in frames, m stands for number of minutes from end to search in.

The algorithm begins by searching for a starting image within the last minute of a recording. The image is found if the similarity exceeds 85 %. Once a match is found, the end image is set as a target and the next 10 seconds of recording are skipped as most of the weather forecast segments are longer, and shorter ones are irrelevant for the dataset. If the image similarity is below 70 %, the algorithm advances by 0.5 seconds. The similarity between 70 % and 85 % means that the algorithm is likely approaching the target image, and the algorithm jumps forward only for 0.25 seconds. If neither the start nor end frames are found by the time the algorithm reaches the end of the recording, it restarts, expanding the search by one more minute. This process continues until both images are found or the entire recording has been searched.

B. Annotation

The target of the second part of the pipeline is to annotate extracted recordings in the most automated way possible. Since the dataset lacks captions, the only way to generate text is to use automatic speech recognition. We used *UWebASR* [11], a web-based automatic speech recognition engine for Czech and Slovak. Except for its web user interface, it provides HTTP API with several examples. It accepts an audio file and outputs the results in multiple formats from plain text with punctations, line breaks, etc. to the web video text tracks format – captions with timestamps and transcript support.

The overall transcription generation process consists of several steps. First, the weather forecast segment is loaded, and audio is extracted. *FFmpeg* was utilized for the audio extraction. Then the *UWebASR* is used to generate the plain text representation of the audio. Finally, the generated text is loaded and split by words. All punctuation is removed except terminal punctuation marks. These were replaced by the special token, which might be useful for context-based machine learning solutions.

IV. Results

A. Preparation pipeline results

The segment extraction part of the preparation pipeline was successfully tested. The expected input, daily news with a weather forecast segment shorter than one minute at the end of the news, took approximately 5 seconds to extract. The algorithm found a target segment in 99 % of cases, the rest was logged and later manually inspected. Some of these were false negatives and small changes of similarity thresholds allowed the algorithm to pass. However, from later testing, it is safe to say that all fault recordings were eliminated during this phase. The outcome of this part was 4706 weather forecast segments, out of the 4713 initial daily news recordings.

The annotation part was done via the *UWebASR* tool. The process took on average around 30 seconds per recording. This part was dependent on 3rd parties' software and could not be further optimized. The transcription was precise, and the overall quality was outstanding, including the punctuation.

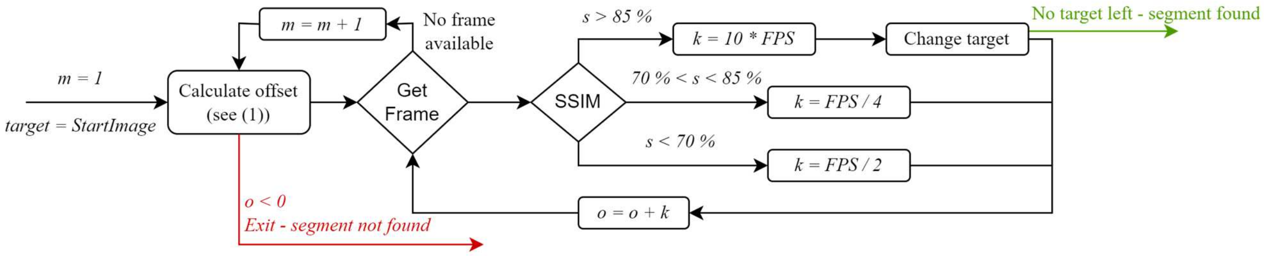

Fig. 4. The logic of segment extraction optimization.

TABLE I. DATASET STATISTICS

Number of samples	4699
Total length of samples (input)	44 h 32 min
Average length of sample (input)	34.1 s
Maximum/minimum length of sample (input)	91.2 / 23 s
Number of speakers (man / woman)	8 (2 / 6)
Average length of sample (output)	57.4 words
Maximum/minimum length of sample (output)	138 / 23 words
Vocabulary size	1679 words

The outcome was manually checked; the audio of 7 recordings was found faulty and was removed. This left the dataset with 4699 recordings. Except for the faulty recordings, there were some verbal slips by the speaker or mistakes on the ASR side. The latter was concentrated in certain patterns or phrases and was manually fixed.

Overall, the pipeline showed promising results decreasing the time that would be required to create and annotate such a dataset manually. Although a manual check is necessary, most of the process is automated and the number of samples to examine is drastically lowered.

B. Dataset results

The outcome of the preparation pipeline is 4699 annotated recordings of weather forecast, all the parameters are shown in Table 2. Although the dataset presented in [8] has over 8000 recordings, they usually consist of one short sentence or phrase. The average length of recoding is 34.11 seconds, while the maximum is 91 seconds, and the minimum is 23 seconds. It was observed that the dataset contains several outliers regarding the length, however their omitting is dependent on the usage. The number of speakers is 8; 6 women and 2 men, majority of them are present across all 14 years of recordings. The dataset was divided into train, test, and validation sets in ratio 80:10:10. The test and validation recordings were selected to ensure an even distribution in the year of their creation. As mentioned earlier, the resolution and overall quality vary across the dataset, and it is up to researchers to adapt it to their algorithms.

From the perspective of annotations, the dataset does not provide gloss annotations, but rather only spoken language representation. The average length of a sequence of words (end of sentence token included) is 57.4. The maximum and the minimum lengths are 138, and 23, respectively. In comparison to [8], the output sequences are much longer, allowing for sign language translation on a scale of several sentences. Most of the output sequences adhere to the following pattern: a quick summary of the weather the next day, temperatures in general and on mountains, and farewell. Sometimes some extra info about the weather situation is added.

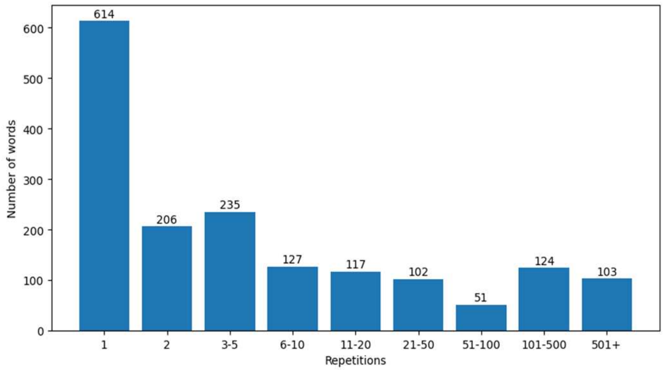

Fig. 5. The graph of repetitions across the vocabulary.

The vocabulary of the dataset is 1679 words, which is smaller compared to [8] 1066 glosses, and 2887 words. Fig. 5 shows the repetition of individual words in the vocabulary. Interestingly, nearly half of the vocabulary is in the dataset only 2 times or less. However, there are 227 words, which are repeated more than one hundred times. The vocabulary size and word repetition align well with the goals of this paper. Both aspects make it suitable for testing and prototyping machine learning solutions for sign language translation. The vocabulary, while relatively small, contains enough frequently repeated words appearing in various contexts.

V. CONCLUSION

The paper shows a successful application of the automation-driven pipeline to prepare a dataset for sign language translation. The approach is divided into two phases: segment extraction and annotation. The former is done via image comparison with SSIM, the latter uses automatic speech recognition *UWebASR*. The method extremely reduces the time, that is required to create such a dataset while minimizing the needed input from a user. The only required input is the video, start image and end image.

The outcome of this paper is not only the pipeline itself but also a demonstration of its usage for creating a Czech Sign Language dataset of weather forecasts. We optimized the method to a point, where the typical recording took less than a minute to process. The output is a dataset consisting of 4699 recordings with a vocabulary of 1679 words. Each recording is annotated with a sequence of words of spoken language. Unlike other datasets for sign language recognition, we omit gloss translations as they proved to be unnecessary for today's natural language processing methods [9].

Apart from this, future work will be focused on developing a machine-learning solution for Czech Sign Language translation. The main priority is to adopt natural language processing methods, which are undergoing swift development and have gained significant popularity recently. Furthermore, we aim to extend this dataset, not only by incorporating additional weather forecast data but also by potentially including other topics if the translation method proves successful.

[1] Krejsa, J. & Vechet, S., "Czech sign language single hand alphabet letters classification," 2020. *19th International Conference on Mechatronics - Mechatronika 2020*, pp.1-5.

[2] Snajder, J. & Bednarik, J., "Czech sign language single hand alphabet classification with Mediapipe," 2022. *International Conference Engineering Mechanics 2022,* pp. 381-384.

[3] Zhang, F. et al., "MediaPipe Hands: On-device real-time hand tracking," 2020. *ArXiv.org*.

[4] Snajder, J. & Krejsa, J., "Classification of Czech sign language alphabet diacritics via LSTM," 2022. *20th Internation Conference of Mechatronics – Mechatronika 2022*, pp. 1-5.

[5] Sincan, O. M. & Keles, H. Y., “AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods,” 2020. *IEEE Access*, *8*, pp. 181340-181355.

[6] Li, D., Rodriguez, C., Yu, X., Li, H., “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison,” 2020. *IEEE/CVF winter conference on applications of computer vision*, pp. 1459-1469.

[7] Šnajder, J. & Krejsa, J., “MediaPipe and its suitability for sign language recognition,” 2023. *29th International Conference Engineering Mechanics 2023*, pp. 251-254.

[8] Camgoz, N. C., Hadfield, S., Koller, O. Hermann, N., Bowden, R., “Neural sign language translation,” 2018. *IEEE conference on computer vision and pattern recognition*, pp. 7784-7793.

[9] Camgoz, N. C., Koller, O., Hadfiel, S., Bowden, R., “Sign language transformers: Joint end-to-end sign language recognition and translation,” 2020. *IEEE/CVF conference on computer vision and pattern recognition*, pp. 10023-10033.

[10] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., “Image quality assessment: from error visibility to structural similarity,” 2004. *IEEE Transactions on Image Processing*, pp. 600–612.

[11] Svec, J., Bulin, M., Prazak, A., Ircing, P., “UWebASR – Web-based ASR engine for Czech and Slovak,” 2018. *CLARIN Annual Conference 2018 Proceedings*, pp. 190-193.