*Proceedings of the 8th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'22)*
*Prague, Czech Republic - July 28- 30, 2022*
*Paper No. MHCI 109*
*DOI: 10.11159/mhci22.109*

# Video Analysis Tool with Template Matching and Audio-Track Processing

**Pragati Chaturvedi, Yasushi Akiyama**
Department of Mathematics and Computing Science, Saint Mary's University
923 Robie Street Halifax NS, Canada
Pragati.Chaturvedi@smu.ca; Yasushi.Akiyama@smu.ca

***Abstract*** - In the last few decades, we have observed the rapid advancement of multimedia analysis tools, and video analysis is one of such application domains. While much effort has been put into the analysis of business and professional videos (e.g., films, professional sports, security cameras) by utilizing advanced image processing algorithms, many of these approaches often do not work well with raw videos that are recorded with a single, consumer-level camera (e.g., a mobile phone) by non-professional videographers. These "amateur" videos typically do not have multiple view-angles and often contain low-resolution and noisy images, making it more difficult to apply certain algorithms compared to cases with videos that are professionally recorded with multiple high-quality cameras and that are properly edited. In this paper, we discuss a prototype interactive video image analysis tool that combines both the image and audio analysis of such videos. The tool provides multiple channels of data analysis visualizations that presumably complement each other for the users to understand the video content effectively and more easily.

***Keywords*:** video/image analysis, audio analysis, multimedia tools, natural language processing (NLP), interaction design

## 1. Introduction

This paper presents a video analysis tool that employs the video audio track analysis (e.g., narrations, commentators) using natural language processing (NLP) as well as the image analysis using template-matching. By visualizing analysis data such as detected keywords and topics (i.e., groups of keywords), Bag of Words (BoW), and the likelihood of video frames containing detected scenes/objects of sports plays, the users of this tool can easily search the video segments of their interests and quickly obtain the data feedback visually. This approach thus likely reduces the amount of time that is typically spent on scanning through the given video to find video segments containing specific actions (e.g., in soccer, when the goalkeeper interacts with the ball, corner kicks, players being subbed, etc.).

In professional sports, the teams, the leagues, and the broadcasters often have substantial resources to set up and control multiple cameras from different angles, and they can hire professional video editors and sound engineers to create special video highlights. In amateur leagues, however, such resources are limited, and videos are usually recorded using a single camera, often with mobile phones or consumer-level camcorders. In some cases, it is difficult to employ common video segmentation techniques (e.g., scene detections based on video cuts [1], or using spectator/audience audio cues [2]). Our approach attempts to overcome some of the shortcomings of these techniques that cannot easily be applied to raw, unedited, single-camera videos. Previously, we have implemented the single-camera sports-video analysis tool based on the template-matching algorithms [3][4], and the current tool presented in this paper is a new extension that adds the video's audio-track analysis based on the NLP techniques to provide the users with another layer of video analysis visualization and interaction. The combination of both the image and audio analyses makes this tool a more powerful and versatile analysis tool as they can complement each other.

## 2. Related Work

While video image analysis for object recognition and automatic video segmentation has progressed greatly in the past few decades, augmenting them with the audio track analysis including transcription and text/language analysis can enhance viewers' understanding of the video content. Using audio cues for video content analysis is not new. For example, Baillie and Jose [2], proposed an audio signal analysis algorithm for the spectators of sports videos to detect certain plays based on Hidden Markov (HMM) model classifiers. The algorithm first segments audio using the BICseg algorithm and the HMM model is trained to classify key events in the video. Another example system is developed by Li and Dorai [5] for

instructional video content analysis. The system first segments based on the audio track into homogeneous audio segments, each of which has a salient sound characteristic such as speech or music, and then extracts discussion scenes in the video based on a Gaussian mixture model (GMM). It then classifies the extracted discussion scenes into either two-speaker or multi-speaker discussions using an adaptive model-based clustering approach. Lao et al. [6] proposed a two-step racquet-hit detection approach to classify events in tennis videos. They achieved the event classification by first mapping the sample-level feature space with the semantic level, and employed heuristic rules based on the knowledge of the sport. Some systems combine both visual and audio analyses to provide a better user experience. Evangelopoulos et al. [7] use saliency in audio based on multifrequency waveform modulations, in the text by tagged keywords in the subtitles of movies, and in video images with a spatiotemporal attention model driven by intensity, colour, and motion. The results of these saliency analyses are displayed as curves so that the users can visualize potentially salient events in the video. Qi et al. [8] developed a news video analysis system by integrating the image and audio analysis results for identifying news segments. Their approach capitalized on the text recognition on video frames and NLP techniques to categorize news stories based on the extracted text. Owens and Efros [9] proposed an approach using a neural network to model how visual and audio components of a video signal are processed as a fused multisensory representation rather than separately. They successfully applied this approach for three areas (1) sound source localization, (2) audio-visual action recognition, and (3) on/offscreen audio source separation.

Our work presented in this paper is the early result of an attempt to increase the usability and user experience of the video analysis tool that provides the visualization of and interactions with the analysis data, capitalizing both on the image and audio-track analyses.

## 3. Tool Overview

This section describes the features of the proposed tool as well as the details of the backend audio and image analysis framework.


Fig. 1: (a) File selection dialogue: the user can either choose the videos that were previously analyzed or upload a new one. (b) Video player view: it provides typical video playback functions (on the left), as well as the video analysis controls (on the right).

### 3.1. Workflow and User-Interactions

To initiate the analysis process, the user first selects a video that they want to analyze. The videos that were previously analyzed by the tool are shown in the *Select a Video* dropdown list (as shown in Fig. 1(a)), and the user can simply choose one from them. Alternately, the user can choose to upload a new video, in which case, a new audio analysis process will automatically begin (as discussed in Section 3.3). When a previously analyzed video is chosen or the audio analysis for a new video is complete, then the user will see the video player view (shown in Fig. 1(b)). In this view, they can perform simple video playback actions (play, pause, fast-forward, rewind, mute/unmute, the caption on/off), as well as interact with the three basic audio analysis features, (a) a keyword search that allows the user to search a specific word, (b) a list of frequently used words, and (c) detected topics that are essentially clusters of words that are related to each other. Any action with these audio analysis features will result in highlighting (in yellow) the video timeline, which indicates the corresponding video segments where the search keyword, the selected frequently used word, or the selected topic was found (shown in Fig. 2). Using these highlighted cues, the user can

easily find the video segments of their interest and further investigate the Bag of Words (BoW) for each of those sections (shown as the pop-up overlay when the mouse cursor hovers over the section).

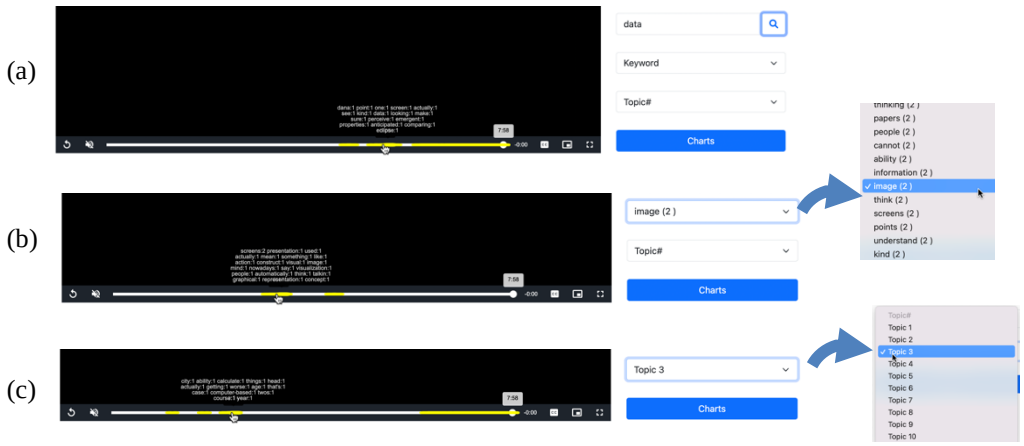


Fig. 2: The three basic audio analysis control panel: (a) the keyword ("data") search results, (b) the list of frequently used words, and (c) the topics (groups of words). When the user searches with a keyword or selects a word or a topic from the lists, the tool highlights the corresponding video segments in yellow to indicate where the selected word/topic appear. The user can also see the BoW (with the frequencies of the words) for each of the highlighted segments.

In addition, the audio analysis pane, which is generated with pyLDAvis tool [10], displays the more detailed audio track analysis results (shown in Fig. 3). To navigate to this view from the main screen, the user clicks on the *Charts* button (Fig. 1(b)). This screen has two parts. On the left is the chart that shows the detected topics (i.e., groups of words) that are projected on the 2D plane based on the topic distances determined by the multidimensional scaling (MDS), and on the right shows the list of words corresponding to the selected topic. The user can directly click on this visualization or use the dropdown list at the top-left to choose a topic. The more detailed discussions of pyLDAvis are provided by the developer of the tool [10].

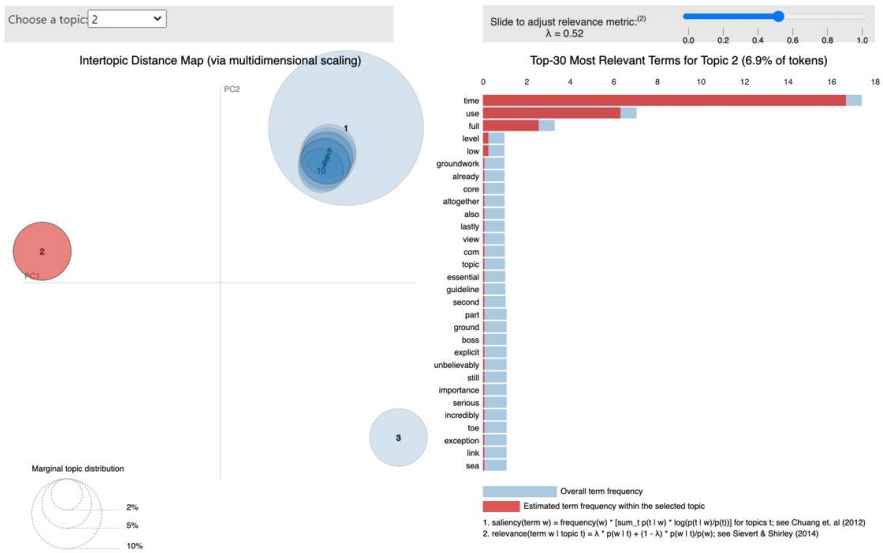


Fig. 3: The visualization pane of the audio analysis results using pyLDAvis [10]. The left-hand side pane shows the detected topics (groups of words). The user hovers over the topic and the corresponding list of the words that appear most frequently on that topic is shown in the right-hand side pane.

The user can also perform the image analysis using template-matching, by selecting a reference area (i.e., a portion of the image that contains an object of interest, such as a soccer goal, as shown in Fig. 4(a)), then the tool will calculate the likelihood of each video frame containing the selected reference area (as discussed in Section 3.2). The results of this image analysis will also be displayed on the video timeline to indicate the likelihood by way of the brightness of the green highlighting (i.e., the brighter the highlighted sections, the more likely that those frames contain the target reference area), shown in Fig. 4(b). The overview of the entire task-flow and the user interactions provided by the proposed video analysis tool is shown in Fig. 5.

(a

(b

Fig. 4: (a) Template matching with the selected reference area. (b) The resulting highlights of the template-matching (in green) and the audio analysis (in yellow), indicating the video segments of interest.

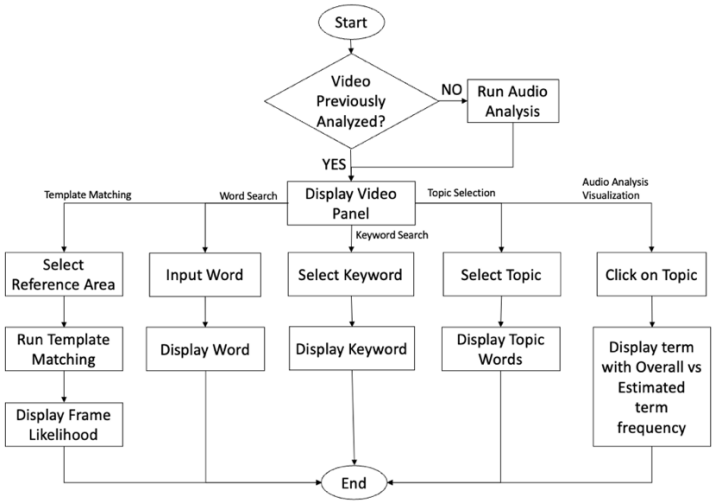


Fig. 5: The overview of the task-flow and the user interactions provided by the proposed video analysis tool. The user can interact with this tool to run both the image and audio analyses and visualize their results to identify interesting sections of the video.

### 3.2. Image Analysis with Template Matching

In the previous version of the tool, we introduced approaches to identifying video frames that potentially contain scenes or objects that the user is interested in (e.g., a soccer goal in a soccer video) by analysing video images based on the template matching, using a simple template-based template matching [3] as well as the three prominent feature-based template matching algorithms [4], namely Oriented FAST and Rotated BRIEF (ORB) [11], Scale Invariant Feature Transform (SIFT) [12], and Speeded Up Robust Feature (SURF) [13]. The tool then displays the likelihood of each video frame showing the object of interest in that frame. More detailed discussions of these algorithms as well as the UI of the original analysis tool can be found in [3][4].

While this approach yielded the reasonable results for the videos that we tested and found video segments that showed soccer plays near the goal, there are cases in which the play was not important or interesting (e.g., the ball was

passed back all the way to the goalkeeper by their teammate, without any opponent players near the goalkeeper). In these cases, it would be ideal if there is additional information available to further filter the template-matching results. This was the motivation to integrate another channel of the video analysis capitalizing on the audio track.

### 3.3. Audio-Track Analysis with Natural Language Processing

The audio track analysis can be divided into three parts, (1) extracting and cleaning the audio track from the video, (2) transcribing the audio to text, and (3) retrieving the desired information from the text. For the extraction of the audio track, the video is converted into an audio file (44.1 kHz, 16-bit mono WAV file) by using FFmpeg [14]. This WAV file will then be read into chunks of milliseconds and the maximum amplitude for each chunk will be obtained. An average of these maximum amplitudes will be calculated and used to eliminate audio segments with their maximum amplitude below this threshold value. The elimination is a two-step process: (i) finding silence gaps longer than two seconds[i], and (ii) marking the starting and ending points of the *active* blocks (i.e., audio segments that have loud enough sound).

The WAV file will then be divided into active blocks and each block will be transcribed using the SpeechRecognition library [15] and saved as a standard .vtt file to be used for the subtitle feature of the video viewer (see Fig. 1(b)). After removing stop-words and punctuations, a Bag of Words (BoW) will also be retrieved for each transcribed block as well as for the entire audio track using the gensim library [16]. This BoW for the entire audio track will be used to populate the dropdown list of the frequently used words in a descending order (see Fig. 2(b)).

The spaCy parser [17] is then used to perform part-of-speech tagging and only nouns, adjectives, and adverbs are kept. Verb forms were removed, and the lemmas of the words are used in the list. Bigrams and trigrams are also built and used to generate a dictionary and term-document frequencies, which will further be passed as a parameter to generate the Latent Dirichlet allocation (LDA) model [18] for potential topics in the transcribed text. These topics are used to populate the topic dropdown list (shown in Fig. 2(c)), and visualized through pyLDAvis tool [10] (shown in Fig. 3). Finally, as discussed earlier, the keyword search as well as the selection of a word/topic will result in the corresponding audio segments to be highlighted (in yellow), and by hovering the mouse cursor over these highlighted section, the users can see the BoW with frequencies to further analyze the words that appear in that segment (shown in Fig. 2).

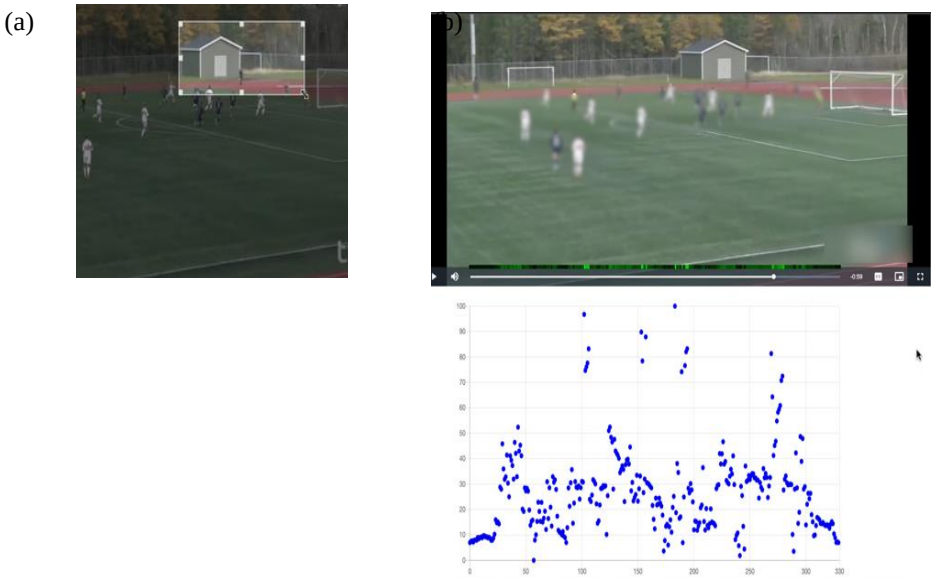


Fig. 6: (a) The reference area used, and (b) the results of the template matching.

[i] While this value is experimentally determined that gave the best performance for our testing purposes and the length of the gap itself does not affect the tool's performance, it should be determined based on the users' preference

## 4. Early Results

In this section, we present and discuss the early sample results of the proposed video analysis tool, tested on a soccer video (5 minutes and 30 seconds long). In this experiment, we tried to identify video segments of the plays in or near the goal area based on the template-matching. We detect the plays only when they happen on the right-hand side the pitch to simplify our discussions here. Fig. 6 shows the result of this image analysis (i.e., the green highlights as the supplementary chart below it), which is the likelihood of the video frames containing the selected template (the the right-hand side). We can further analyse these segments with the brighter green (i.e., higher likelihood) together the audio-track analysis. For example, when you explore a few different topics, we can see that some topics seem related to the brighter green segments than the other topics do (e.g., Topic #4 segments highlighted in yellow has more correlated segments with the green highlights than Topic #2 segments do, as shown in Fig. 7).


Fig. 7: Comparisons of (a) Topic #2 segments versus (b) Topic #4 segments against the image analysis results (green highlights).

Another observation is that the selection of the word “ball” (which was mentioned 11 times in this short video) from the list also seems to align roughly with the green highlights (shown in Fig. 8). This is not surprising because, when the ball is near the goal/goalkeeper, it is likely mentioned more frequently than other occasions.

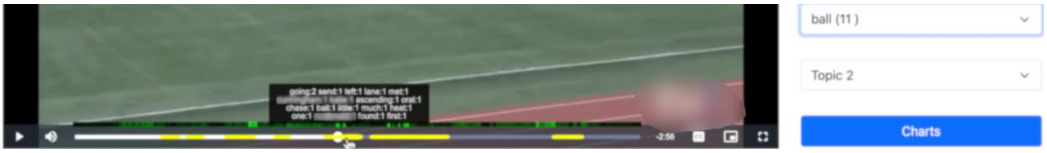


Fig. 8: The highlighted segments (in yellow) that contain the word "ball" compared to the template-matching results (in green).

Finally, an interesting and important observation is when we select the name of the goalkeeper from the word list (the name is blurred in the figure to protect their privacy), we can see in Fig. 9 that not all the yellow segments align with the green highlights. A further investigation revealed that the goalkeeper’s name was mentioned when they actually interacted with the ball (e.g., shots on target but saved by the goalkeeper), but for other times, although the template-matching found the target object in some of the video frames, the play itself may not have been as interesting/important (e.g., the goalkeeper did not come near the ball), thus this extra (audio) layer of the video analysis was useful in filtering some sections that may not be as exciting as the image analysis alone suggested.

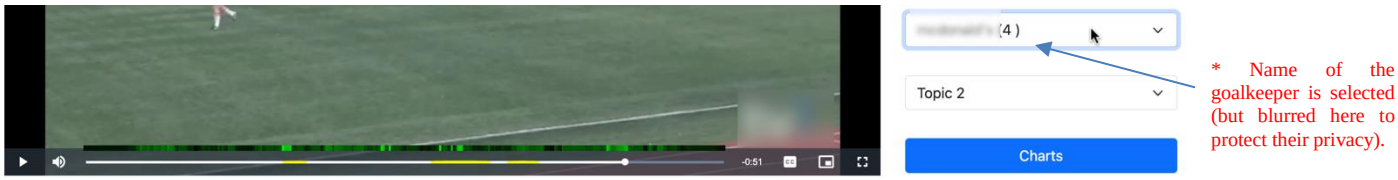


Fig. 9: The highlighted segments (in yellow) that contain the goalkeeper's name (blurred to protect their privacy).

## 5. Conclusion and Future Work

In this paper, we presented a video analysis tool that provides both the image and audio-track analyses and visualizes the resulting data for the users to view and interact with them. The image analysis is completed based on the template-matching algorithms while the video's audio-track is transcribed and segmented for several natural language processes. By adding the audio-track analysis layer to the existing template-matching approaches, the resulting visualization allows the users to easily locate the video segments, especially when one mode of the analysis is insufficient to identify interesting segments and filter out unwanted ones.

We are currently preparing for a set of user-studies to evaluate the usability of this tool. Based on the outcome of these studies, we will iterate through the common design cycles to refine the tool's interface design. We are also exploring the application domains of this tool beyond the sports videos. For example, given the recent shift to online delivery of classes at educational institutions worldwide, we now have an abundance of lecture videos. Our tool may be used to analyze these videos (e.g., scan the lecture slides and analyse the instructor's spoken words), and allow the students to understand and navigate through their content more easily without manually scrubbing the video. Another potential domain is for creating labelled data for training object recognition models. By highlighting only relevant video segments, it can likely save the annotators substantial time for searching for target objects in a long video when labelling objects.

## 6. Acknowledgements

We would like to thank Ms. Hemanchal Joshi, who has implemented the feature-based template matching algorithms in the backend, and Mr. Gurpartap Singh, who has contributed his time to the development of a part of the UI for the template-matching analysis visualization. We would also like to thank Atlantic University Sport (AUS) for granting us the permission to use their videos for our research.

## 7. References

[1] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," IEEE Transactions on Image Processing, vol. 12, no. 7, pp. 796–807, Jul. 2003.

[2] M. Baillie and J. M. Jose, "An audio-based sports video segmentation and event detection algorithm," in 2004 Conference on Computer Vision and Pattern Recognition Workshop, Jun. 2004, pp. 110–110.

[3] Y. Akiyama, R. Garcia, and T. Hynes, "Video scene extraction tool for soccer goalkeeper performance data analysis," in Proceedings of ACM IUI 2019 Workshop UISTDA2019 User Interfaces for Spatial and Temporal Data Analysis, Los Angeles, USA, Mar. 2019.

[4] H. Joshi, "Image feature matching for scene extraction from single-camera videos," MSc thesis, Saint Mary's University, 2021.

[5] Y. Li and C. Dorai, "Instructional video content analysis using audio information," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 6, pp. 2264–2274, 2006. doi: 10.1109/TASL.2006.872602.

[6] W. Lao, J. Han, and P. H. N. de With, "Automatic sports video analysis using audio clues and context knowledge," in EuroIMSA, 2006.

[7] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis, "Video event detection and summarization using audio, visual and text saliency," 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3553–3556, 2009.

[8] W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang, "Integrating visual, audio and text analysis for news video," in Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101), vol. 3, 2000, 520–523 vol.3. doi: 10.1109/ICIP.2000.899482.

[9] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," CoRR, vol. abs/1804.03641, 2018.

[10] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 63–70.

[11] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to sift or surf," in 2011 International Conference on Computer Vision, 2011, pp. 2564–2571. doi: 10.1109/ICCV.2011.6126544.
[12] D. Lowe, "Object recognition from local scale-invariant features," in Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, 1999, 1150–1157 vol.2. doi: 10.1109/ICCV.1999.790410.
[13] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in Computer Vision – ECCV 2006, A. Leonardis, H. Bischof, and A. Pinz, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 404–417, isbn: 978-3-540-33833-8.
[14] S. Tomar, "Converting video formats with FFmpeg," Linux Journal, vol. 2006, no. 146, p. 10, 2006.
[15] A. Zhang, Speech Recognition (version 3.8) [software], Available from https://github.com/Uberi/speech recognition, 2017.
[16] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, vol. 3, no. 2, 2011.
[17] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017.
[18] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of Population Structure Using Multilocus Genotype Data," Genetics, vol. 155, no. 2, pp. 945–959, Jun. 2000.