Mehmet Akif Ersoy Üniversitesi Uygulamalı Bilimler Dergisi
Journal of Applied Sciences of Mehmet Akif Ersoy University

Design of Audio Description System Using Cloud Based Computer Vision

Emre KARAGÖZ[1*], Kutan KORUYAN[2],

[1]Dr., Dokuz Eylül University, Distance Education Application and Research Center, İzmir, Turkey

[2]Asst. Prof. Dr., Dokuz Eylül University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, İzmir, Turkey

Geliş Tarihi/Received: 26.11.2019
Kabul Tarihi/Accepted: 01.02.2020

Doi: doi.org/10.31200/makuubd.651261
Araştırma Makalesi/Research Article

ABSTRACT

Developments and changes in multimedia tools are actively used in many areas of life and bring a huge value to them. Nowadays, the concept of artificial intelligence is highly developed and there are hundreds of practices and methods to support the living standards especially for people with disabilities. The system developed in this study enables automatic visualization of the media output scenes such as movies, documentaries, etc., which are visually impaired people by means of computer vision technique, and the results are transferred to the users by voice command. HTML5 and CSS are used for visualizing the system, PHP and JAVASCRIPT are used for programming. MySQL is preferred as the database of the system. Computer vision, translation from text to speech and translation from one language to another are the main instruments used in this study. Cloud-based Microsoft AZURE Computer Vision API is used for computer vision, Javascript Responce.js library is used for text-to-speech translation, Google Cloud Text-To-Speech and Microsoft Azure Text to Speech APIs are used for translation from one language to another one.

Keywords: Audio Description, Computer Vision, Text to Speech Translation, Machine Translation, Cloud Computing.

* Sorumlu yazar/Corresponding author
E-mail/e-ileti: emre.karagoz@deu.edu.tr

Bulut Tabanlı Bilgisayarlı Görü Kullanılarak Sesli Betimleme Sistem Tasarımı

ÖZET

Multimedya araçlarındaki gelişim ve değişimler hayatın birçok alanında aktif şekilde kullanılmakta ve büyük oranda artı değer kazandırmaktadır. Yapay zekâ kavramının son derece gelişmiş olduğu günümüzde, özellikle engelli bireylerin yaşam standartlarını destekleyecek yüzlerce uygulama ve metot bulunmaktadır. Bu çalışmada geliştirilen sistem özellikle görme engelli bireylerin izledikleri film, belgesel gibi video formatındaki medya çıktı sahnelerinin görüntü imgeleme tekniği sayesinde otomatik olarak betimlenmesini ve sonuçların kullanıcılara sesli olarak aktarılmasını sağlamaktadır. Sistemin görselleştirilmesinde HTML5 ve CSS, programlanmasında PHP ve JAVASCRIPT dilleri kullanılmıştır. Sistemin veritabanı olarak MySQL tercih edilmiştir. Yapay zekâ ve bilişim teknolojilerinden olan bilgisayarlı görü, metinden konuşmaya çevirme ve bir dilden başka bir dile çeviri, bu çalışmada kullanılan temel enstrümanlardır. Görüntü imgeleme işlemleri için bulut tabanlı Microsoft AZURE Computer Vision API, metinden sese çevirme için Javascript Responce.js kütüphanesi, bir dilden başka bir dile çeviri işlemlerinde ise Google Cloud Text-To-Speech ve Microsoft Azure Text to Speech API'leri kullanılmıştır.

Anahtar kelimeler: Sesli Betimleme, Bilgisayarlı Görü, Metinden Konuşmaya Çeviri, Makina Çevirisi, Bulut Bilişim.

1. INTRODUCTION

Media with visual features such as videos, documentaries and films can become meaningful for visually impaired individuals with the help of dubbing. This method, called audio description (AD), is the technique used for making theatre, movies and TV programmes accessible to blind and visually impaired people: an additional narration describes the action, body language, facial expressions, scenery and costumes (Benecke, 2004: 78). As a vital service for all visually impaired people, AD is used in cinemas, playfields and museums (Whitehead, 2005: 962). AD is a labour-intensive action based on certain rules, where the descriptor must be trained and voiced by experts in his/her field. In addition, the number of people working for audio description of a 2-hour movie can be up to 60 people (Lakritz and Salway, 2006: 2). There are some guidelines for AD that require professional talent. The American Council of the Blind and the Lifelong Learning Program under the European Union developed several guidelines (ADI AD Guidelines Committee, 2003; Remael et al, n.d.). In addition, Netflix, which includes AD in some of its publications, also has its own AD manual (Netflix, 2019).

A number of studies have been carried out to facilitate the process of AD, which is a costly, labour intensive and professional process, to increase the accessibility of visually depicted media content for visually impaired individuals. Lakritz and Salway (2006) showed that they could create AD text semi-automatically from the film scenarios and their work could help create AD text. Delgado et al. (2015) presented a semi-automated AD system using speech recognition, machine translation and text-to-speech (TtS) translation. In the study of Jang et al. (2014), they presented a semi-automatic descriptive video service by using the audio and subtitles in the videos and placing AD text in these spaces. Gagnon et al. (2010) accelerated the process of video annotation using computer vision (CV) technologies for voice and face recognition and developed an application that can automatically detect many elements of video annotation. In their study, Rohrbach et al. (2017) explained exactly what is in the image in AD texts and that AD texts are more suitable to train possible software that recognizes the objects in the images.

These studies are applications that will provide a source of data for voice over AD, text creation or machine learning (ML). The main point of this study is the description of the actions and elements of AD using CV which is a branch of artificial intelligence (AI).

This study has been developed especially for the use of visually impaired individuals. The system, which was created by using CV, depicts the related image automatically in the specified time intervals by means of AI, and the actions in the video are presented to the users through TtS. Since the texts returned from CV system are in English, they are also translated into Turkish for Turkish speaking individuals.

2. COMPUTER VISION

Today, keeping up with the endless speed of technology requires a great effort for users. Especially AI applications bring many advantages and offer solutions for different kind of problems. In addition to these solutions, AI's ability to mimic human intelligence is able to create added value for production with less human labour in production processes. AI is used in many fields such as banking, insurance, security, health, military, natural language processing and image processing (Aydemir, 2018: 29).

CV, a branch of AI, is used to identify objects that the computer sees by imitating human vision. CV is generally defined as a scientific field that extracts information from digital images, algorithms that can understand the content of images and use them for other applications (Krishna, 2017: 17). Thanks to CV, images and videos can be analysed automatically by computers (Dawson-Howe, 2014: 1).

CV is an interdisciplinary concept of AI, ML, robotics and geometry. These systems enable the renewal of production processes, reduce production costs, increase product quality and guarantee human safety (Klancnik et al., 2015: 571).

Different algorithms are used to perform CV operations. Commonly used algorithms are Viola / Jones Algorithm, Feature-Checking Algorithm, The Detection / Rejection Classifier Algorithm and Artificial Neural Network Algorithms.

In CV processes, MATLAB Computer Vision Toolbox, Microsoft Azure Computer Vision (MACV) and OpenCV tools are used as a software and platforms.

3. TEXT TO SPEECH TRANSLATION

TtS is a computer-based system where the input is text and the output are a simulated vocalization of that text (Pagani, 2005: 957). TtS conversion transforms linguistic information stored as data or text into speech. It is widely used in audio reading devices for blind people (O'Malley, 1990: 17).

A TtS system is a two-step process in which text is converted to its equivalent digital audio. A text and linguistic analysis module process the input text to generate its phonetic equivalent and performs linguistic analysis to determine the prosodic characteristics of the text. A waveform generator then produces the synthesized speech (Carvalho et al., 1998: 1). Nabiyev (2016: 699) divided the transformation from text to speech synthetically into two parts:

- Human read and record words: Requires large memory usage. Restricted words reduce the success of the system and hinder the flexibility of the system. For end-added languages such as Turkish, this approach remains simple.
- Reading and recording syllables by human: It is formed by combining letter sounds in appropriate tone according to the location of the letters forming syllables.

While softwares for TtS was used in the past, there are now TtS systems that serve as cloud-based Application Programming Interface (API). Examples of these are Google Cloud Text-To-Speech, IBM Watson Text to Speech, Microsoft Azure Text to Speech (MATtS), and Amazon Polly.

4. MACHINE TRANSLATION

Machine translation (MT) is the automatic translation of texts from one language to another with computers and mobile devices without human intervention (Aslan, 2018: 89). The MT system has a large database and has a wide range of applications including information communication in foreign languages, fast access to information in multilingual international organizations, translation of foreign trade documents and synchronous translation (Nabiyev, 2016: 477). MT is one of the subfields of automatic translation and it is a field of AI and computational linguistics. Today, Google Cloud Translate (GCT) and Microsoft Azure Translator Text (MATT) are the most popular MT applications.

GCT enables enterprises to dynamically translate between languages using Google's pre-trained or custom ML models based on your content needs. With AutoML Translation, it allows developers with limited ML expertise to train high-quality custom models (Google Cloud, n.d.).

MATT API supports text translation across more than 60 supported languages in mobile, desktop and web applications through the open REST interface (Microsoft Azure, n.d.).

5. SYSTEM DESIGN

This study is a prototype study based on the process of analysing video frames per second of some sample videos with CV. The aim of the study is to help visual depictions of visually impaired individuals while watching visual media. A 64-bit server based on Linux is used to make the system work, Javascript and PHP are used as system programming languages. The main reason for choosing these languages is to establish a web-based structure that will enable easy integration and accessibility of the system to each location. MySQL is the preferred database. When the system is taken as a whole, the processes are shown in Figure 1.

The steps from the management area to the user area are as shown in Figure 1. Firstly, the selected video is uploaded to the server and the pictures are stored in the system by taking one frame from each video (Figure 1, (1)). Each of these image files is called the time of the frame in the video to which it relates. For example, the name of the image in the 10th second is set to 10.jpeg. Then, each of these photos is sent to the MACV API (Figure 1, (2)), the result from the API (with the help of AI, description of actions in the images) and the confidence (with the help of AI, the score between 0 and 1, indicating the accuracy of the description of actions in the images) are added in related tables in the database (Figure 1, (3)). After this operation, the ID numbers are identified by matching the MACV data stored in the database with the pictures. The results from MACV are in English and each text in the database is translated from English to Turkish with the help of MATT and GCT APIs and the texts are added to the related fields in the database (Figure 1, (4)). The purpose of using both the MATT and GCT APIs is to measure the accuracy of both machine converter APIs. Table 1 shows the example data in the database.

After all these operations are done, the user can watch the videos and the relevant data is taken from the database at the time intervals defined by the system administrator (eg. 1 frame / 5 sec) and transmitted to the user via MATtS (Figure 1, (5)). In addition, the MACV result of the relevant scene in the area at the bottom of the screen is displayed in Turkish and English languages. The MACV API delivers results at a certain confidence level. For example, if the MACV output for a scene yields a confidence level of 0.90, it is highly probable that the objects on the scene are accurately predicted although in some scenes is estimated as 0.30, 0.40 by MACV. The administrator is able to ignore the underlying database data by setting a specific threshold level. That is, if a threshold of 0.85 is set, lower rate responses in the database are ignored. This enables the user to transmit more successful and more accurate answers. The system currently has 10 videos for users. Figure 2 shows the user interface.

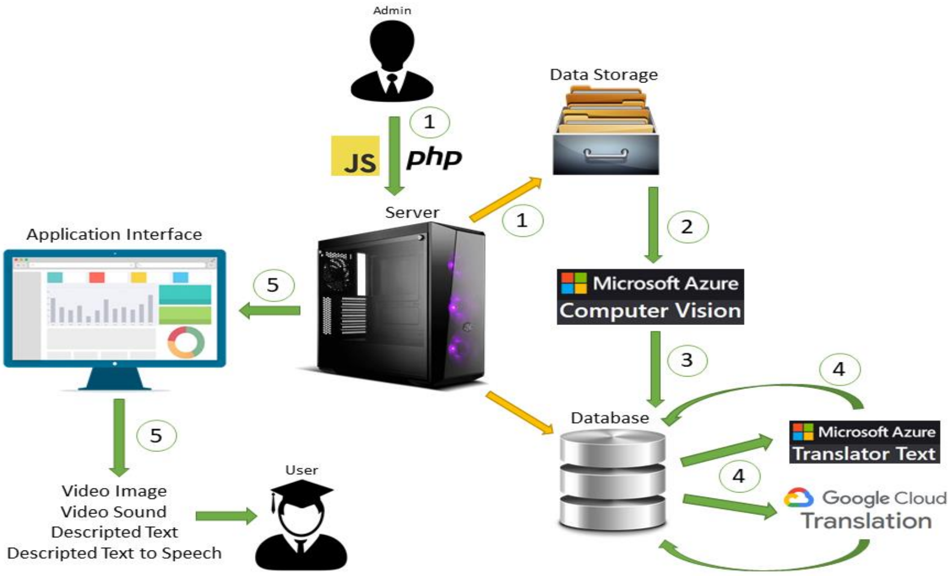
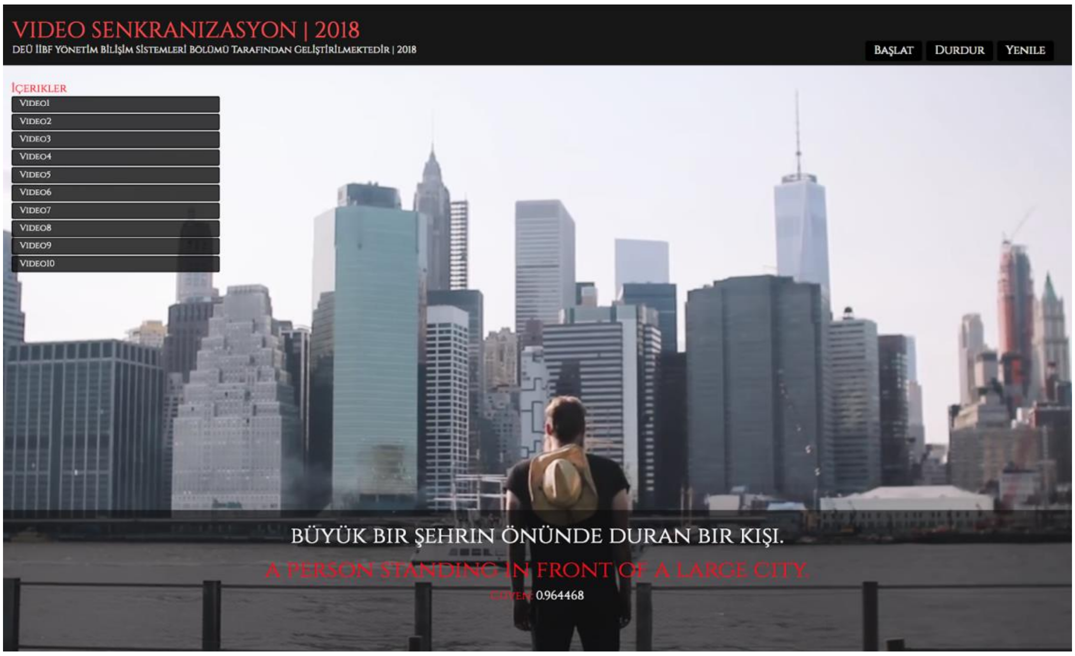


Figure 1. System working process: (1) The video is captured every second and stored on the server. (2) These images are then sent to MACV API. (3) MACV responses (results and confidence) are added to the database. (4) Each MACV printout stored in the database is translated from English to Turkish by MATT and GCT. (5) While the user watches the videos, he/she pulls the relevant data from the database and transfers it to the user by voice.

Table 1. Example data from the database

id	Frame	Content	Confidence	GCT Results	MATT Results
1	0	a group of clouds in the dark.	0.603847	-	-
2	1	a group of clouds in the sky.	0.85317	gökyüzünde bulutlar bir grup.	gökyüzünde bir grup bulutlar.
3	2	a castle on a cloudy day.	0.813688	bulutlu bir günde bir kale.	bulutlu bir gün bir kale.
4	3	a castle on a cloudy day.	0.823829	bulutlu bir günde bir kale.	bulutlu bir gün bir kale.
5	4	-	0	-	-
6	5	a group of clouds in the sky.	0.877947	gökyüzünde bulutlar bir grup.	gökyüzünde bir grup bulutlar.
7	6	a castle on a cloudy day.	0.790941	bulutlu bir günde bir kale.	bulutlu bir gün bir kale.
8	7	a group of people walking down a street.	0.95886	bir grup insan bir sokakta yürürken.	bir grup insan sokakta yürüyor.
9	8	a group of people walking down the street.	0.982836	bir grup insan sokakta yürürken.	sokakta yürüyen bir grup insan.
10	9	a group of people walking down a street.	0.955556	bir grup insan bir sokakta yürürken.	bir grup insan sokakta yürüyor.
11	10	a group of people walking down the street.	0.961385	bir grup insan sokakta yürürken.	sokakta yürüyen bir grup insan.



VIDEO SENKRANIZASYON | 2018

DEÜ İİBF YÖNETİM BİLİŞİM SİSTEMLERİ BÖLÜMÜ TARAFINDAN GELİŞTİRİLMEKTEDİR | 2018

BÜYÜK BIR ŞEHRIN ÖNÜNDE DURAN BIR KIŞI.

A PERSON STANDING IN FRONT OF A LARGE CITY.

GÜVEN: 0.964468

Figure 2. User interface

6. RESULTS

This study is based on the answers given by MACV, MATT and GCT APIs. These responses were examined one by one by the system administrators at 5 second intervals and their accuracy were determined. One author made an optimistic and more tolerable assessment while the other author made a pessimistic assessment. Table 2 shows optimistic evaluation results, while Table 3 shows pessimistic evaluation results.

Table 2. Optimistic result table

Video ID	Total Frame	MACV Correct	MACV Wrong	MACV Correct Wrong Ratio	GCT Correct	GCT Wrong	GCT Correct Wrong Ratio	MATT Correct	MATT Wrong	MATT Correct Wrong Ratio
Video1	91	60	31	1.94	71	20	3.55	78	13	6
Video2	90	71	19	3.74	83	7	11.86	89	1	89
Video3	76	63	13	4.85	55	21	2.62	70	6	11.67
Video4	112	92	20	4.6	84	28	3	106	6	17.67
Video5	137	99	38	2.61	110	27	4.07	122	15	8.13
Video6	51	37	14	2.64	38	13	2.92	45	6	7.5
Video7	176	135	41	3.29	171	5	34.2	171	5	34.2
Video8	121	80	41	1.95	104	17	6.12	114	7	16.29
Video9	48	34	14	2.43	43	5	8.6	46	2	23
Video10	87	70	17	4.12	73	14	5.21	60	27	2.22
Total	989	741	248		832	157		901	88	
%		74.92	25.08		84.13	15.87		91.1	8.9	

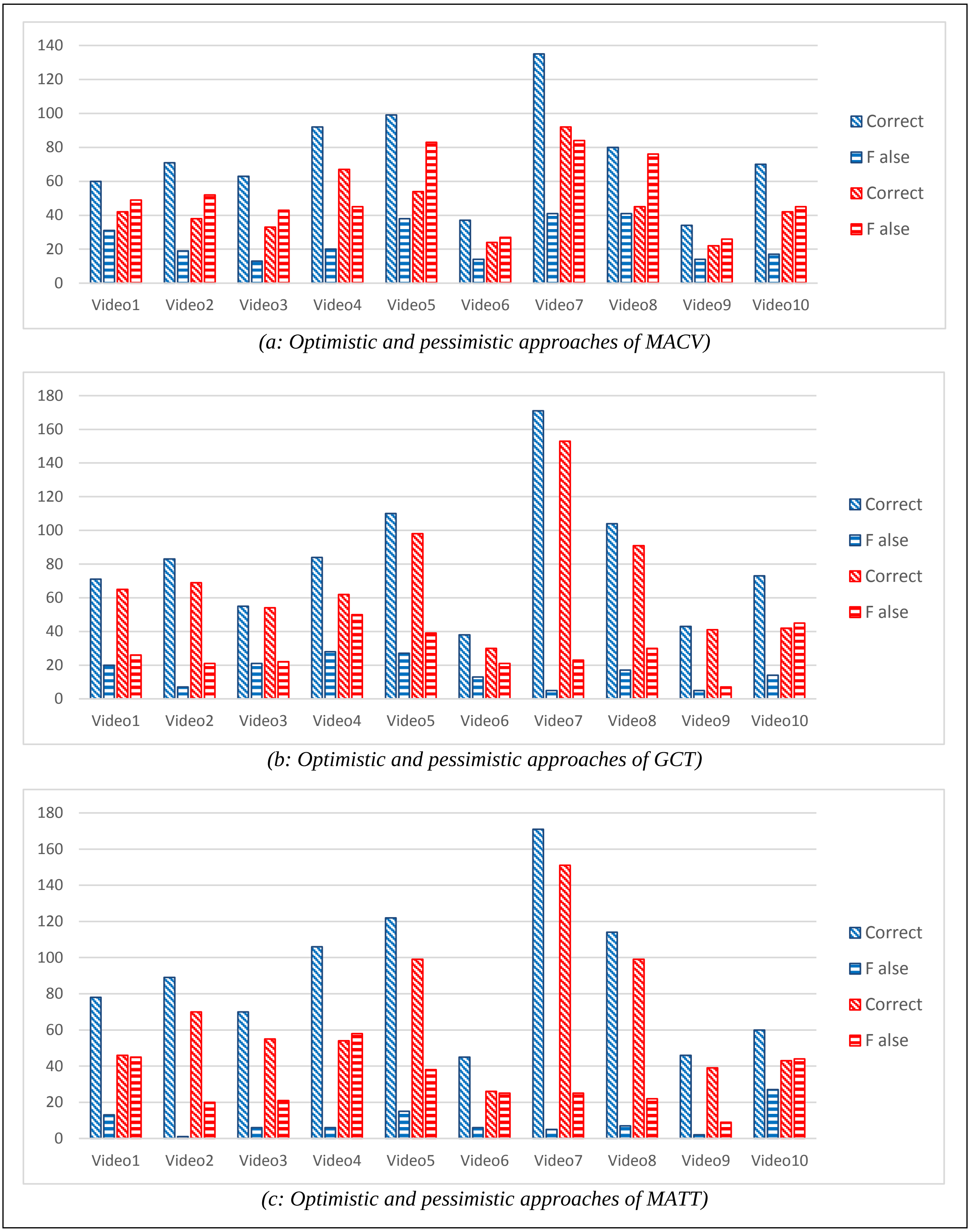
As shown in Table 2, 91 scenes were selected for video 1, and with an optimistic assessment of MACV API responses, 60 were accepted to be correct and 31 were wrong. Confidence levels of MACV responses were ignored when performing these evaluations. Therefore, MACV outputs that are considered wrong by the evaluator seem likely to have a low confidence level. Likewise, for the 1st video, 71 of the GCT API's English to Turkish translations were considered correct, 20 were wrong, and 78 of the MATT API translations were considered correct and 13 were false. The MACV API accuracy for all videos is 75.92%. When GCT and MATT were compared, the translation rates into Turkish were 84.13% and 91.1%, respectively.

Table 3. Pessimistic result table

Video ID	Total Frame	MACV Correct	MACV Wrong	MACV Correct Wrong Ratio	GCT Correct	GCT Wrong	GCT Correct Wrong Ratio	MATT Correct	MATT Wrong	MATT Correct Wrong Ratio
Video1	91	42	49	0.86	65	26	2.5	46	45	1.02
Video2	90	38	52	0.73	69	21	3.29	70	20	3.5
Video3	76	33	43	0.77	54	22	2.45	55	21	2.62
Video4	112	67	45	1.49	62	50	1.24	54	58	0.93
Video5	137	54	83	0.65	98	39	2.51	99	38	2.61
Video6	51	24	27	0.89	30	21	1.43	26	25	1.04
Video7	176	92	84	1.1	153	23	6.65	151	25	6.04
Video8	121	45	76	0.59	91	30	3.03	99	22	4.5
Video9	48	22	26	0.85	41	7	5.86	39	9	4.33
Video10	87	42	45	0.93	42	45	0.93	43	44	0.98
Total	989	459	530		705	284		682	307	
%		46.41	53.59		71.28	28.72		68.96	31.04	

According to the pessimistic assessment, Table 3, 42 of the MACV results of 91 scenes for the 1st video was accepted as correct and 49 of them were wrong. For the same video, 65 of the GCT results were considered correct, 26 of them were incorrect, and 46 of the MATT results were considered correct, and 45 were considered incorrect.

For all videos, the accuracy of the MACV API is 46.41% according to the pessimistic evaluation. When GCT and MATT were compared, the conversion rates into Turkish were 71.28% and 68.96%, respectively. Figure 3a (MACV), 3b (GCT) and 3c (MATT) are graphical representations of the results produced by optimistic and pessimistic approaches.



(a: Optimistic and pessimistic approaches of MACV)

(b: Optimistic and pessimistic approaches of GCT)

(c: Optimistic and pessimistic approaches of MATT)

Figure 3. Graphical representations of the results produced by optimistic (blue) and pessimistic (red) approaches

7. CONCLUSIONS AND RECOMMENDATIONS

This study has been developed to provide support for visually impaired individuals to follow visual media such as video. The analyses were performed on the basis of MACV and the authors attempted to evaluate them subjectively. MACV yields 74.92% optimistic and 46.41% correct results in 10 videos. Therefore, obtaining the results of the other CV libraries on the same data and then comparing them with the MACV data in the study is a matter that should be discussed in future studies. According to obtained results, the library with the best results should be used.

Since the results from MACV are in English, texts for Turkish speaking users have been translated into Turkish with the help of GCT and MATT APIs. Accuracy rate of GCT with optimistic and pessimistic approach is 84.13% and 71.28%, respectively. The accuracy of MATT with optimistic and pessimistic approach is 91.10% and 68.96%, respectively. According to these results, it is seen that MATT is more successful in machine translation in 989 sentences.

Image data added to the database were obtained at a 1 second interval. However, in one video, different scene transitions can occur within 1 second and this is causing errors in the transfer of results to the user. It is thought that it can be a more positive and correct method to determine these stage transitions with another algorithm and to realize information transfer while the user is watching video.

10 videos were used in the study and these results were evaluated and tried to test the success. The integration of more videos into the system can help to achieve more accurate results. In addition, more people performing video evaluations will be useful in measuring the effectiveness of the system.

The system is thought to provide a separate data source for the AD text writers and can be used as a semi-automatic AD text generator.

It is thought that this method can be used for different purposes such as selecting the English equivalents of objects on the screen in fields such as foreign language education.

REFERENCES / KAYNAKLAR

ADI AD Guidelines Committee. (2003). Guidelines for Audio Description, Retrieved: 23.06.2019, from http://www.acb.org/adp/guidelines.html

Aslan, E. (2018). Otomatik çeviri araçlarının yabancı dil öğretiminde kullanımı: Google çeviri örneği. *Selçuk Üniversitesi Edebiyat Fakültesi Dergisi*, *0*(39), 87-104.

Aydemir, E. (2018). *Weka ile yapay zeka*. Ankara: Seçkin Yayıncılık.

Benecke, B. (2004). Audio-description. *Meta*, *49*(1), 78-80.

Carvalho, P., Trancoso, I. M. & Oliveira, L. C. (1998). Automatic Segment Alignment for Concatenative Speech Synthesis in Portuguese, *Proc. of the 10th Portuguese Conference on Pattern Recognition, RECPAD'98*, Lisbon.

Dawson-Howe, K. (2014). A practical ıntroduction to computer vision with opencv. ABD: John Wiley & Sons.

Delgado, H., Matamala, A. & Serrano, J. (2015). Speaker diarization and speech recognition in the semi-automatization of audio description: An exploratory study on future possibilities? *Cadernos de Tradução*, *35*(2), 308-324.

Gagnon, L., Chapdelaine, C., Byrns, D., Foucher, S., Heritier, M. & Gupt, V. (2010). A computer-vision-assisted system for Videodescription scripting. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, San Francisco, CA, USA, 1-8.

Google Cloud. (n.d.). *Cloud translation*. Retrieved: 15.05.2019, https://cloud.google.com/translate/

Jang, I., Ahn, C. & Jang, Y. (2014). Semi-automatic DVS authoring method. *Computers Helping People with Special Needs: 14th International Conference, ICCHP 2014*, Springer International Publishing, Switzerland.

Klancnik, S., Ficko, M., Balic, J. & Pahole, I. (2015). Computer vision-based approach to end mill tool monitoring. *International Journal of Simulation Modelling*, *14*(4), 571-583.

Krishna, R. (2017). *Computer vision: Foundations and applications*. Stanford: Stanford University.

Lakritz, J. & Salway, A. (2006). *The semi-automatic generation of audio description from screenplays*, Dept. of Computing Technical Report, University of Surrey, UK.

Microsoft Azure. (n.d.). *Translator text API*. Retrieved: 26.06.2019, https://azure.microsoft.com/en-gb/services/cognitive-services/translator-text-api

Nabiyev, V.V. (2016). *Yapay zeka (5th Ed.).* Ankara: Seçkin Yayıncılık.

Netflix. *Netflix audio description style guide v2.1*. Retrieved: 12.11.2019, https://partnerhelp.netflixstudios.com/hc/en-us/articles/215510667-Audio-Description-Style-Guide-v2-1.

O'Malley, M. H. (1990). Text-to-speech conversion technology. *Computer*, *23*(8), 17-23.

Pagani, M. (2005). *Encyclopedia of multimedia technology and networking*. Hershey PA, USA: Idea Group Inc.

Remael, A., Reviers, N. & Vercauteren, G. (n.d.). *ADLAB Audio Description guideline*. Retrieved: 24.06.2019, http://www.adlabproject.eu/Docs/adlab%20book/index.html.

Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A. & Schiele, B. (2017). Movie description. *International Journal of Computer Vision*, *123*(1), 94-120.

Whitehead, J. (2015). What is audio description. *International Congress Series*, *1282*, 960-963.