Museum Guidance in Sign Language: the SignGuide project

D.I. Kosmopoulos
C. Constantinopoulos, M. Trigka
D. Papazachariou, K. Antzakas
V. Lampropoulou
{dkosmo,kkonstantino}@upatras.gr
trigka@ceid.upatras.gr
{papaz,k.antzakas}@upatras.gr
V.Lampropoulou@upatras.gr
University of Patras
Patras, Greece

A. Argyros, I. Oikonomidis
A. Roussos, N. Partarakis
G. Papagiannakis
{argyros,oikonom}@ics.forth.gr
{troussos,partarak,papagian}@ics.forth.gr
papagian@ics.forth.gr
Foundation for Research and Technology - Hellas
Heraklion, Greece

K. Grigoriadis
kgrigor@mls.gr
MLS Innovation Inc
Thessaloniki, Greece

A. Koukouvou, A. Moneda
{akoukouvou,amoneda}@culture.gr
Archaeological Museum of Thessaloniki
Thessaloniki, Greece

ABSTRACT

We present an overview of the SignGuide project. Its main goal is to develop a prototype interactive museum guide system for deaf visitors using mobile devices that will be able to receive visitors' questions in their native (sign language) with regard to the exhibits and to provide additional content also in sign language using an avatar or video, utilizing techniques from the field of computer vision and machine learning. The paper presents the basic ideas and technologies involved as well as some preliminary results.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Animation**; **Tracking**.

KEYWORDS

museum guidance, sign language translation, video retrieval

ACM Reference Format:
D.I. Kosmopoulos, C. Constantinopoulos, M. Trigka, D. Papazachariou, K. Antzakas, V. Lampropoulou, A. Argyros, I. Oikonomidis, A. Roussos, N. Partarakis, G. Papagiannakis, K. Grigoriadis, and A. Koukouvou, A. Moneda. 2022. Museum Guidance in Sign Language: the SignGuide project . In *PETRAE '22: PErvasive Technologies Related to Assistive Environments, June 29– July 1, 2022, Corfu, GR.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3529190.3534718

1 INTRODUCTION

Sign Languages (SLs) are the main means of communication for deaf people. The access to SL is essential for the fulfillment of basic Human Rights. However there is a shortage of interpreters, which undermines these rights. One of those rights is the access to cultural and educational content, like the content presented in museums.

The SignGuide project aims to develop a prototype interactive museum guide system for deaf visitors using mobile devices that will be able to acquire videos with visitors' questions in their native SL with regard to the exhibits and to provide additional content also in sign language using an avatar or video.

Probably, vision is the only sensor modality that could be of practical use because (a) only vision can capture manual and non-manual cues, which provide essential information for the SL, (b) camera-equipped hand-held devices with powerful processors are a widely available and (c) recent advances in computer vision and machine learning render SL translation a realistic option.

The SignGuide project aims to fulfill the following goals:

- Implementation of a mobile service that recognizes the user's query in SL about the exhibition
- Implementation of a service that analyzes the previous query and retrieves additional content in SL relevant to the visitor query.
- Development of an avatar for the presentation of the additional material through SL.
- Development of accompanying material in such a way as to support retrieval based on free text queries

The long-term viability of the suggested application will be achieved by (a) simple off-the-shelf equipment for users, (b) efficient implementation of the proposed algorithms, (c) simple installation, and (d) development of a business plan to enable the proposed system's lifespan.

The consortium is composed of (a) The Signal Processing and Telecommunications Lab of the University of Patras as experts in machine learning, which is necessary for the recognition of GSL, and the Deaf Studies Unit at the same university, who will bring the users, the interpreters and the GSL experts, (b) the Computational Vision and Robotics Lab of ICS-FORTH, which will adapt their 3D hand model for tracking, and the Human Computer Interaction Lab

PETRAE '22, June 29–July 2, 2022, Corfu, GR

ACM ISBN 978-1-4503-9631-8/22/06…$15.00
https://doi.org/10.1145/3529190.3534718

of the same institute, who are developing the avatar, (c) the MLS Innovation Inc, who specialize in language technologies products, and (d) the Archaeological Museum of Thessaloniki, who will install the system in their exhibition.

2 THE SYSTEM

Here, we present some related works, intended contributions and goals for the development of the proposed system.

Related work: The importance of interacting with deaf users in their native SL has been recognized by museums, and is mainly served by interpreters[1]. Applications that provide recorded content as short SL videos have been commercialized or explored and supported by a rather small number of museums[2]. Events for the deaf are organized at the British Museum and National Gallery, and multimedia guides are provided in the British SL[3,4]. The Vatican Museum offers scheduled tours of Italian SL twice a week, and there is also a possibility of multimedia tours using portable guides in American SL[5]. The Smithsonian American Art Museum organizes guided tours for deaf visitors, and the presentations have been videotaped[6]. Also, little research has been done internationally to close the loop of interaction between native SL users and online agents, and that is where the SignGuide interactive platform aims. Although museum stands or electronic guides via mobile devices support several languages for user interaction, SL is not included. The use of SL for interaction with electronic guides (a) significantly increases usability in people with severe communication difficulties and (b) does not disturb the museum environment in contrast to spoken language.

Innovation: For the first time in a museum environment, tools will be integrated that allow the complete communication of the deaf speakers of SL, closing the loop of interaction in their native language. SignGuide, unlike the usual approaches, proposes a system that allows the SL speakers to formulate their questions in SL. The system presents additional multimedia material related to SL speakers' queries in SL too. The application is based on the latest developments in the field of machine learning and computational vision for the analysis of gestures and the dynamic synthesis of content. Relevant combined approaches are virtually non-existent internationally.

Proposal: The platform includes the interconnected software, tools, units (see Figure 1) that will be installed in the Archaeological Museum of Thessaloniki:

(1) Optical analysis tool for Sign-Language queries.
(2) Content Retrieval Tool, based on SL queries and will be combined with the knowledge base, from which it will draw content based on queries in SL. In addition, the knowledge base will communicate with third-party service providers and will receive the relevant data.
(3) User interface on a "smart" device with the Android operating system. It will receive and send image-video-text data to the other subsystems and present the most relevant results.
(4) Virtual guide unit, integrated into the user interface, whose main functions are the conversion of text to SL (Text-to-SL) as well as modelling, drawing performance and the final rendering of 3D graphics (avatar).
(5) Exhibit tracking tool. It works with visual detection to correlate material search with the exhibit. Developed in the MuseLearn project and will be adapted [33] to SignGuide, contributing to more natural user interaction.

The envisaged use scenario has as follows. The SL user enters the museum and installs the application on his/her mobile phone via a wireless network or is provided with a tablet with the application installed by the museum. In the beginning, he/she optionally enters some demographics for statistical purposes. Then, the user targets the exhibit of interest with the camera of the mobile device. The exhibit is recognized by the system using artificial vision and the user is given the choice to enter a query about the exhibit, the room, the time period, etc., in natural SL. The phrase is identified and semantically analyzed. The system presents the video/virtual agent to SL retrieving the relevant information and composing the relevant material. The user's interest is extracted based on his/her feedback (number of clicks, content viewing time, etc.).

3 SUBSYSTEMS

In the following, we are going to present our approach to some of the key subsystems that we aim to develop.

3.1 Human tracking and SL translation

Related work on visual monitoring of user's body: The problems of visual localization and monitoring of human body posture (hands, face, torso) are directly related to the understanding of SL. Such methods have attracted the interest of researchers in recent years [14], [26], [31], [27], [18], [3],[6], [38], [37]. Methods that assess the posture of the human hand can be categorized into discriminative and generative, as described, for example, in the publication [31]. The first category includes methods that learn a direct mapping from the visual data into the pose space and apply this mapping directly to the visual input. While they off-load costly calculations this way, their disadvantage is the limited estimation accuracy. In the second class of methods, i.e., the generative ones, the geometric model of the human hand is used as a structural element to formulate a problem of optimizing the parameters of the human hand pose. This approach usually has increased computational costs during execution compared to discreet methods but, on the other hand, allows the improvement of the hand pose estimation accuracy depending on the computational resources devoted to the solution of the optimization problem. Hybrid methods combine features of both the above categories.

Despite the recent advancements, e.g., [31], [27], [18], [3] crucial improvements are still needed. The main problem to be faced by methods applied to monitor gesturing hands is the speed and flexibility with which a hand can move [14]. Further problems include A) the initialization of the monitoring process as well as the re-initialization in case of failure of the monitoring, and B) the

[1] https://www.metmuseum.org/events/programs/access/visitors-who-are-deaf
[2] https://signly.co/
[3] https://www.britishmuseum.org/learning/access.aspx
[4] https://www.nationalgallery.org.uk/visiting/access/deaf-hard-of-hearing
[5] http://mv.vatican.va/3_EN/pages/z-Info/Didattica/MV_Info_Didattica_05_sordi.html
[6] http://americanart.si.edu/education/asl/

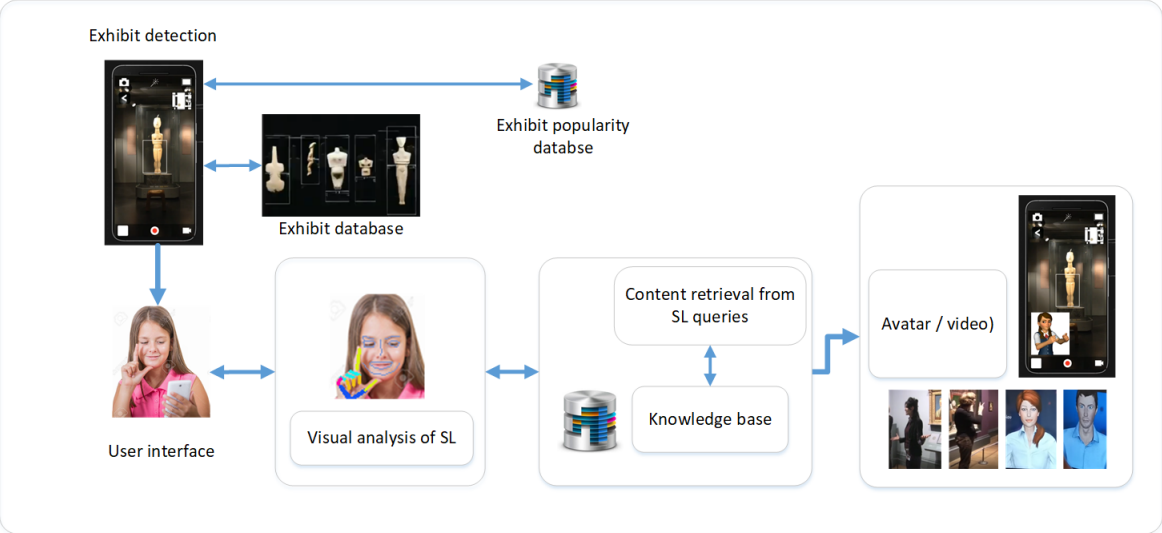

Figure 1: The architecture of the proposed SignGuide platform. The visitor sees the exhibit through a mobile device. The exhibit is then located by the system (based on the image) and the user queries in SL. After analyzing the query, the appropriate digital material is synthesized and presented through a virtual guide. At the same time, the users' choices are recorded based on visiting for further analysis.

adaptation of the used model of appearance and hand kinematics to the respective user.

Related work on Modeling SL: The works that achieved the best results recently were based mainly on two databases, SIGNUM (laboratory) [1] and PHOENIX [21] (weather reports in SL). However, there is no database for queries related to museum tours in SL, and it needs to be developed. In the problem of SL analysis, the ability of the neural networks was partially utilized. In [23], a categorization training was proposed using weakly characterized sequential data, by integrating a convolutional network (CNN) into an iterative EM algorithm, which, in combination with [22], significantly reduced the error. The paper [4] introduced a deep learning architecture, where the problem is decomposed into a subsystem sequence, the SubUNets. As input handheld images, the network first extracts spatial information via a CNN network and instead of HMM (as in previous works, e.g. [8]), a two-directional LSTM is introduced. [9] proposed a deep architecture with a three-step optimization process for weakly characterized data. The utility of (two-directional) LSTM networks was also exploited in [24], with the integration of a CNN-LSTM network in combination with a repetitively trained HMM model. As a continuation of the same technique in [25], more extensive experiments are performed, comparing different CNN network architectures. The state-of-the-art is by using a winner-takes-all and transformer networks [36].

Despite their usefulness, the above techniques are directly dependent on high-dimensional holistic data without taking into account the inherent limitations of the language or the physiology of the human body. Also, the training from SL data, given SL's "richness", creates a high need for a large volume of data. Hence, data-intensive approaches like ones mentioned above face difficulties in real settings.

Innovation: We suggest estimating and monitoring the 3D hand and body pose to identify (and synthesize) Sign Language queries related to a museum tour. We suggest the use of deep learning convolutional neural networks, taking into account possible hand-body layouts and linguistic limitations. We also aim to reduce the dimension of the problem by modelling the different channels of information (much lower dimension) separately with fusion then.

Approach: With the initial goal of hand monitoring for the recognition of SL gestures to museum visitors, we will proceed to the improvement of the hand monitoring method [27], which uses a hybrid approach, to specifically meet the requirements of this application. In particular, an effort will be made to improve the accuracy of the estimation, suitable for estimating SL gestures. Peculiarities of the approach to a museum are (a) the environment that will not be controlled because a museum is a public place (moving visual background, lighting) and (b) the possible need to model the SL with one hand since the other hand is most likely engaged in holding the device (c) extracting meaning from phrases rather than individual words. In addition, aiming at synthesizing GSL phrases with a virtual character, we will develop and apply a method of assessing a human body pose using a colour image as input. That will allow the extraction of information that can be used to compose phrases from the proposed virtual character.

More specifically, we will record sequences from hand, body, and face pose for use as input to the learning processes. Entering the process of locating hands and monitoring their pose are sequences of color images as captured by a color camera. Optionally, we can use a depth rating sensor (RGB-D camera) or stereoscopic images for training. In the advance (off-line) process, we will record predetermined gestures and estimate the observed poses using the

available visual information (color, depth or stereoscopic pair), taking into account the possible hand/body poses. In the next step, the visual information sequences along with poses will be used as input data for the learning process to specialize the point estimate process in the most common poses used in SL.

As part of the improved hand tracking method [27], the accuracy of hand observation can be improved by using depth data, using larger spatial and/or temporal analysis data, and using stereoscopic images. In the first direction, we can experiment with data from an RGB-D sensor, allowing us to train our models with the method of weak supervision. In particular, the availability of in-depth scene information can be used during the training, even if it is not available on the device when the trained model is used to assess a hand pose in the museum environment. This can be achieved by using in-depth information as an additional network learning goal during training and then ignoring this result at the estimation stage. (Depth information is critical for the isolation of the visual background, while input normalizations are expected to compensate for changes in lighting). Relevant works [3] have shown that this strategy can improve the pose estimation accuracy, as, indirectly, the network is forced to learn the 3D structure of the observed hands.

For large-scale SL modelling, it has been shown that the use of linguistic constraints such as hand layout, trajectory, and orientation can give good results [10]. Also, characteristics of the body or facial expression play an important role in semantics. We will try to integrate the relevant constraints in our model, such as Bayes prior probability, which is expected to bring the model closer to the optimal solution.

Also, the holistic representation of the features via convolutional networks is not necessarily the optimal one since the useful visual information concerns many independent channels both in the hands (pose, distance from the torso) and the face (expressions with the eyes, eyebrows and the mouth). Considering them as a single visual input through a monolithic vector of features extracted from a convolutional network facilitates the need to solve a high dimension problem, which generally requires a huge volume of relevant data, which is difficult to collect and annotate. Our approach aims to capitalize upon recent advancements (e.g., [36]) by treating separately different information channels using RNNs or sequence-to-sequence models (see Fig. 2).

3.2 Content retrieval based on queries in SL

Related work: Dynamic content retrieval for museum tours has not been sufficiently utilized. Much more, this applies to SL. For example, the Boston Museum of Fine Arts [15] has developed American Sign Language videos, which are used to guide 53 collections and 145 points and are statically associated with unique exhibits. However, systems for finding specific content in video files have been developed for many years. SQL languages can be used to express the requested content to be retrieved from a video library. A mechanism for composing relevant snippets from video files is used to present the results [17]. Similar systems have been proposed, such as the FrameQL query language used to identify desired video clips using machine learning techniques [19] and situation locating [2]. It seems that there is no comprehensive support system for dynamic content retrieval in museum environments for SL-supported tours. There are general-purpose museum item management systems, several static content systems, and far fewer and fragmented attempts at semantic query systems to locate the desired content.

Innovation: The proposed system will have the following innovations: (a) Organized multimedia content structure, which will enable integration in the result of retrieving SL video explanatory files or the ability to integrate avatar. (b) A query mechanism for locating suitable material from the database, composing and reproducing the material in an appropriate sequential format. (c) Investigation of methods of answering queries using machine learning.

Approach: The information about the exhibits will be stored in a database, while the exhibits' information will be interconnected with the video files or corresponding texts that will be articulated by an avatar in SL. When the visitor moves into the exhibition, depending on which exhibit his/her camera is pointing at, on the screen of his/her device the appropriate material is displayed accompanied by a video in SL or avatar. An important parameter in the above structure is the interactive nature of the narrative, as the visitor, depending on the way he chooses to move in the exhibition, shapes the evolution of the narrative.

Emphasis will be placed on how to build content in a MySQL database, to support fast navigation and search on it as well as linking to relevant multimedia material. For this purpose, the technology for the MAIC trading platform developed by MLS[7] will be used, which will be adapted to the structure of SL. Using open-source internet technologies, the content will be accessed and retrieved from the database. Each piece of information will be mapped to a separate video file in SL or text to be spoken by the virtual guide. The particularity of the implementation is the fast stapling of answers in a single narration so that the visitor begins to see the video/avatar immediately and does not notice either the loading of the individual pieces or any delay or abrupt change during playback. That is, queries to the database will be made based on the search criteria and utilizing the properties of each exhibit and the accompanying material. Based on the results of the queries, the appropriate video files will be identified, the composition of which will be the complete narration for the visitor. Special emphasis will be given to the distribution of content and the processing that will be required for the production of multimedia presentations. More specifically, complex computing will be performed on the server and, the results will be transmitted over the network to the mobile device. Besides, in the user's device, the framework of the application will be stored/installed, but only the content that the visitor recalls will be transferred to it. The presentations will be adapted to the size of the device (responsive design).

Based on user requirements we will seek to support the following semantic axes: Types of information: interpretive, explanatory, ability to observe detail, correlation with other objects in the exhibition or objects in other museums, placement in the original historical and useful context and environment, original form etc. Ways of connecting: alternative scenarios of visiting and linking multimedia information based on (a) thematic tours, based on one or more common features, such as historical period, type of objects, activity (e.g., daily life at home, social life, politics) (b) stories and narratives of historical or fictional figures about aspects of

[7] https://maic.gr/

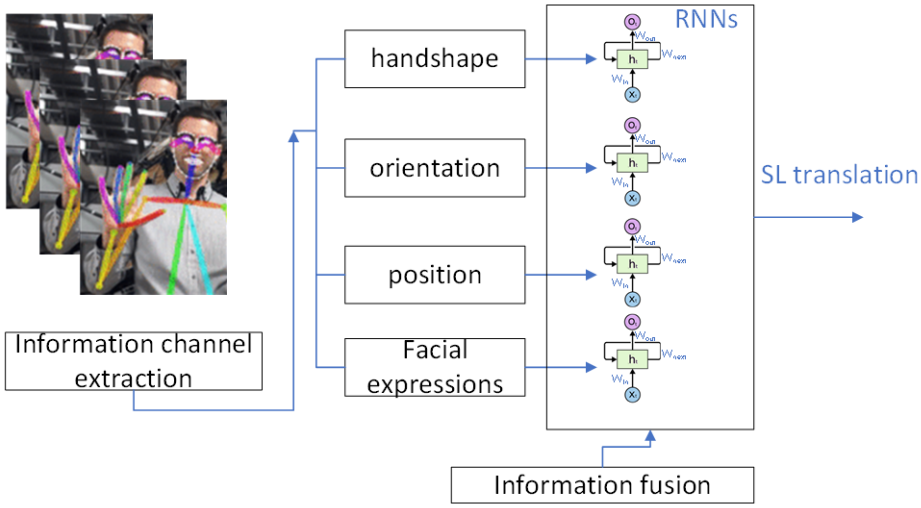

Figure 2: The extraction of skeletal information makes it possible to separately model much simpler different channels (shape, orientation, position, expressions) and then fuse them.

everyday life or important historical events (c) educational games, such as a lost treasure game where museum exhibits become key points of search. Use of standards such as CIDOC [11] for content presentation and further dissemination. Ways of presentation: automatically, by entering an area or focusing on an exhibit, after selection, e.g. menu interaction, activation based on the visitor's profile of interest, activation under conditions, if, e.g., a visitor stays in a place for a long time.

Finally, we will seek to improve/enhance the MAIC platform according to the above axes through the integration of natural language processing methods [32], with the aim of pre-automating the organization process (offline). We will explore chatbot architectures, e.g., ChatterBot[8] and libraries like Natural Language ToolKit[9] to develop generative models that will be able to learn from users' queries without being rule-based. We will explore the possibility of a better question-answer correlation in real-time (online) through the submission of clarifying questions with active learning techniques, e.g., [7].

3.3 Virtual guide in SL

Related work: Virtual characters can be used for interaction through dialogue and gestures due to their innate ability to simulate verbal and non-verbal communication behavior (body motion and gestures). The use of virtual characters as personal and reliable dialogue systems - and especially in the area of sign language that lacks the audio information channel - poses a number of challenges, as it requires not only reliable and consistent movement and dialogue, but mainly non-verbal communication combined with coding emotional components in the way in which the entry into motion is translated by the virtual character.

In addition to modelling logic and creating intelligent behavior, which is an open field of Artificial Intelligence [20], the visual representation of a character includes his/her perceived behavior from a decoding perspective, such as facial expressions and gestures, and it also involves many open issues regarding physical communication [30]. In recent years, geometric algebra with the "Euclidean", "Homogeneous" and "Conformal" models has attracted research interest, since such models can produce more efficient and smooth results than other algebras. Geometric Algebra is a mathematical framework that provides a simple and convenient mathematical model for representing the orientations and rotations of objects in three dimensions and offers the ability to formulate compact transformation algorithms. Conformal Geometric Algebra (CGA) extends the usefulness of 3D Geometric Algebra by extending the category of runners to include translations and expansions (uniform scaling). The rotors are simpler to operate than the Euler angles, more numerically stable and more efficient than the rotation panels for synthesizing transformations, avoiding the "Gimbal Lock" problem. Some CGA applications [12] involve rotation interference, retrospective ray detection for illumination, rotor rotation control and intersection control for collision detection and realistic shadow imaging.

Recently, extensive research has been conducted on the automation of SL movement so that it can be used by virtual characters. The ViSiCast and eSIGN projects [13] focused on developing an infrastructure for translating the text into semantic symbols. Using the text written in English as input, the system translated into written text in Sign Language (English-ESL). This kind of translation, however, is very literal and can cause unnatural results. Written sign language can be translated into symbols using the HamNoSys mark [16], which describes the symbols as positions and movements of both the hands and the rest of the body, i.e., the upper torso, head and face. This notation enables virtual playback, and

[8] ,https://github.com/gunthercox/ChatterBot
[9] https://www.nltk.org/

Figure 3: Avatar-guide that responds to user queries in SL.

the motion system is more flexible and reusable compared to others that use motion data or handmade animation.

Virtual Characters as interlocutors in Virtual and Augmented Reality environments begin to display impressive human abilities of natural dialogue. Virtual Characters of this kind have already been used successfully in various fields. Especially over the last decade, there has been significant progress, including Ada and Grace, a pair of virtual museum guides at the Boston Science Museum [34]. Also quite successful are the INOTS and ELITE training systems at the Naval Station in Newport and Fort Benning [5].

Innovation: Creating virtual characters to present dialogues in sign language has special requirements regarding the realism of the illustration. In addition, due to the lack of audio information channel, emphasis should be placed on proper movement performance, which includes proper hand form performance through the movement of the virtual character-signer's hands in space, as well as synchronization with posture (or movement) of the body and head, and/or facial expression. The combination of the above goes beyond the existing imaging techniques in various fields, such as for example the synchronization of lips with the speech that has specialized characteristics in relation to the traditional approaches. All in all, sign language is a very demanding field in terms of its realistic performance using virtual characters.

Proposal: We will focus research on the realistic simulation of virtual characters in real-time, through the realistic depiction of both the character and his movement (Figure 3). We will incorporate a variety of features such as movement, object handling, eye-focus, lip synchronization to represent lip expressions when signing, etc. The characters will be able to adopt a realistic anthropomorphic look or a cartoon appearance depending on the user group or the preferences of each user or the general characteristics of the user group. For this purpose, a real-time modular virtual character rendering programming framework will be implemented, which includes advanced imaging algorithms (separable subsurface scattering, ambient occlusion, image-based lighting and diffuse shadow mapping, etc.). For this purpose, a 3D game creation machine (Unity3D) will be used for export to multiple computing platforms. In addition, we will extend virtual character lighting models using the GPU [35] to be described using a more general framework of Geometric Algebra, [29], [28]. Autodesk's most popular 3D Studio Max (or 3Ds max) software will be used to design the 3D character (virtual avatar), covering the needs for realistic design, with the addition of lights and shadows. The character and its 3D environment (3D MODELLING) will be designed, while then the skeleton for its movement (rigging) will be created. Rigging is a difficult and demanding technique, and the frame that is made during it (i.e., the rig) must be subjected to frequent tests to check its proper operation on the 3D final model. In the next stage, the skinning process takes place, during which we apply and "tie" the skeleton and the bones that we have made with the skin and the structure of the 3D model of the character.

To complete the design of the 3D character, the following steps should be done (indicatively):

(1) mapping on the 2D surface of the model UVs (uv mapping),
(2) rendering texture and surface material of the 3D model (texturing),
(3) directing and composing a plan (staging),
(4) 3D animation, and
(5) lighting - creating 3D effects.

A graphics rendering engine such as OpenSceneGraph or Unity 3D will be used for real-time graphics performance.

4 SYSTEM EVALUATION

The whole system's success will be based on the evaluation of its subsystems' performance, assuming some criteria, the main of which are presented next. The first criterion will concern the exhibit localization by measuring the deviation from the actual position. Another one will assess the visitor's query recognition (i.e., the creation of targeted answers to questions, the response time, the retraining time of the language recognition model, the visitor's grade, the recovery and communication mechanisms with users, in case of failures, the ease of use and learning of the system, e.t.c.). The avatar will be evaluated based on user-friendliness and rendering time.

5 CONCLUSIONS

We have introduced the basic concepts behind the SignGuide project. Also, the basic elements of the proposed architecture and the principles of their implementation have been presented. In the near future, we are going to begin the implementation and the experimentation with real users. There are a lot of challenges to deal with, the most obvious being the modeling and translation of the continuous SL in real-time and the rendering of the SL into an avatar.

ACKNOWLEDGMENTS

This work is partially supported by the Greek Secretariat for Research and Innovation and the EU, Project SignGuide: Automated Museum Guidance using Sign Language T2EDK-00982 within the framework of "Competitiveness, Entrepreneurship and Innovation" (EPAnEK) Operational Programme 2014-2020.

REFERENCES

[1] Ulrich von Agris and Karl-Friedrich Kraiss. 2010. SIGNUM Database: Video Corpus for Signer-Independent Continuous Sign Language Recognition. In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Philippe Dreuw, Eleni Efthimiou, Thomas Hanke, Trevor Johnston, Gregorio Martínez Ruiz, and Adam Schembri (Eds.). Valletta, Malta, 243–246.

[2] Manish Annappa, Sharma Chakravarthy, and Vassilis Athitsos. 2016. Preprocessing of video streams for extracting queryable representation of its contents. In *International Symposium on Visual Computing*. Springer, 301–311.

[3] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. 2018. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 666–682.

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 3075–3084. https://doi.org/10.1109/ICCV.2017.332

[5] Julia C Campbell, Matthew Jensen Hays, Mark Core, Mike Birch, Matt Bosack, and Richard E Clark. 2011. Interpersonal and leadership skills: using virtual humans to teach new officers. In *Proc. of Interservice/Industry Training, Simulation, and Education Conference, Paper*, Vol. 11358. Citeseer.

[6] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. 2021. Active learning for bayesian 3d hand pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3419–3428.

[7] Sotirios P Chatzis and Dimitrios Kosmopoulos. 2012. Visual Workflow Recognition Using a Variational Bayesian Treatment of Multistream Fused Hidden Markov Models. *IEEE transactions on circuits and systems for video technology* 22, 7 (2012), 1076–1086.

[8] Sotirios P Chatzis, Dimitrios I Kosmopoulos, and Theodora A Varvarigou. 2008. Robust sequential data modeling using an outlier tolerant hidden Markov model. *IEEE transactions on pattern analysis and machine intelligence* 31, 9 (2008), 1657–1669.

[9] Runpeng Cui, Hu Liu, and Changshui Zhang. 2017. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7361–7369.

[10] Mark Dilsizian, Polina Yanovich, Shu Wang, Carol Neidle, and Dimitris Metaxas. 2014. A New Framework for Sign Language Recognition based on 3D Handshape Identification and Linguistic Modeling. In *9th International Conference on Language Resources and Evaluation (LREC 2014)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (Eds.). Reykjavik, Iceland, 1924–1929.

[11] Martin Doerr, George Bruseker, Chryssoula Bekiari, Christian Emil Orey, Thanasis Velios, and Stephen Stead. 2020. *Definition of the CIDOC Conceptual Reference Model Version 6.2.9*. Technical Report. ICOM/CIDOC CRM Special Interest Group.

[12] Leo Dorst, Daniel Fontijne, and Stephen Mann. 2007. *Geometric Algebra for Computer Science (Revised Edition)*. Elsevier.

[13] Ralph Elliott, John RW Glauert, JR Kennaway, Ian Marshall, and Eva Safar. 2008. Linguistic modelling and language-processing technologies for Avatar-based sign language presentation. *Universal Access in the Information Society* 6, 4 (2008), 375–391.

[14] Ali Erol, George Bebis, Mircea Nicolescu, Richard D Boyle, and Xander Twombly. 2007. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding* 108, 1-2 (2007), 52–73.

[15] Hannah Goodwin. 2013. American sign language and audio description on the mobile guide at the museum of fine arts, Boston. *Curator: The Museum Journal* 3, 56 (2013), 369–370.

[16] Thomas Hanke. 2004. HamNoSys-representing sign language data in language resources and language processing contexts. In *LREC*, Vol. 4. 1–6.

[17] Eenjun Hwang and VS Subrahmanian. 1996. Querying video libraries. *journal of visual communication and image representation* 7, 1 (1996), 44–60.

[18] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. 2018. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 118–134.

[19] Daniel Kang, Peter Bailis, and Matei Zaharia. 2019. Challenges and Opportunities in DNN-Based Video Analytics: A Demonstration of the BlazeIt Video Query Engine.. In *CIDR*.

[20] Z. Kasap and N. Magnenat-Thalmann. 2007. Intelligent virtual humans with autonomy and personality: State-of-the-art. *Intelligent Decision Technologies* 1, 1-2 (2007), 3–15.

[21] Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding* 141 (Dec. 2015), 108–125.

[22] Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding* 141 (2015), 108–125.

[23] Oscar Koller, Hermann Ney, and Richard Bowden. 2016. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3793–3802.

[24] Oscar Koller, Sepehr Zargaran, and Hermann Ney. 2017. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4297–4305.

[25] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. 2018. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. *International Journal of Computer Vision* 126, 12 (2018), 1311–1325.

[26] I. Oikonomidis, N. Kyriazis, and A. Argyros. 2011. Efficient model-based 3D tracking of hand articulations using Kinect.. In *BmVC*, Vol. 1. 3.

[27] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. 2018. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 436–445.

[28] Margarita Papaefthymiou, Dietmar Hildenbrand, and George Papagiannakis. 2016. An inclusive Conformal Geometric Algebra GPU animation interpolation and deformation algorithm. *The Visual Computer* 32, 6 (2016), 751–759.

[29] George Papagiannakis. 2013. Geometric algebra rotors for skinned character animation blending. In *SIGGRAPH Asia 2013 Technical Briefs*. 1–6.

[30] P Papanikolaou and G Papagiannakis. 2015. Real-time separable subsurface scattering for animated virtual characters. In *GPU Computing and Applications*. Springer, 53–67.

[31] T. Sharp, D. Keskin, C.and Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, et al. 2015. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 3633–3642.

[32] Roberta Akemi Sinoara, João Antunes, and Solange Oliveira Rezende. 2017. Text mining and semantics: a systematic mapping study. *Journal of the Brazilian Computer Society* 23, 1 (2017), 1–20.

[33] G. Styliaras, C. Constantinopoulos, P. Panteleris, D. Michel, N. Pantzou, K. Papavasileiou, K. Tzortzi, A. Argyros, and D. Kosmopoulos. 2020. The MuseLearn Platform: Personalized Content for Museum Visitors Assisted by Vision-Based Recognition and 3D Pose Estimation of Exhibits. In *Artificial Intelligence Applications and Innovations*, I. Maglogiannis, L. Iliadis, and E. Pimenidis (Eds.). Springer International Publishing, 439–451.

[34] William Swartout, David Traum, Ron Artstein, Dan Noren, Paul Debevec, Kerry Bronnenkant, Josh Williams, Anton Leuski, Shrikanth Narayanan, Diane Piepol, et al. 2010. Virtual museum guides demonstration. In *2010 IEEE Spoken Language Technology Workshop*. IEEE, 163–164.

[35] Mike Tato, Petros Papanikolaou, and George Papagiannakis. 2012. From real to virtual rapid architectural prototyping. In *Euro-Mediterranean Conference*. Springer, 505–512.

[36] A. Voskou, K. P. Panousis, D. Kosmopoulos, D. N. Metaxas, and S. Chatzis. 2021. Stochastic Transformer Networks with Linear Competing Units: Application to end-to-end SL Translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11946–11955.

[37] Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2020. Deep learning-based human pose estimation: A survey. *arXiv preprint arXiv:2012.13392* (2020).

[38] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. 2018. State of the art on monocular 3D face reconstruction, tracking, and applications. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 523–550.