NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

GONÇALO VINAGRE MARTINS

Master in Computer Science

SLVIDEO: A SIGN LANGUAGE VIDEO MOMENT RETRIEVAL FRAMEWORK

MASTER IN COMPUTER SCIENCE AND ENGINEERING

NOVA University Lisbon
September, 2024

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF
COMPUTER SCIENCE

SLVIDEO: A SIGN LANGUAGE VIDEO MOMENT RETRIEVAL FRAMEWORK

GONÇALO VINAGRE MARTINS

Master in Computer Science

Adviser: João Miguel da Costa Magalhães
Associate Professor, NOVA University Lisbon

Co-adviser: Sofia Carmen Faria Maia Cavaco
Associate Professor, NOVA University Lisbon

Examination Committee

Chair: Nuno Manuel Ribeiro Preguiça
Full Professor, FCT-NOVA

Rapporteur: Rui Manuel Feliciano de Jesus
Associate Professor, Instituto Politécnico de Lisboa do Instituto Superior de Engenharia de Lisboa

Adviser: João Miguel da Costa Magalhães
Associate Professor, FCT-NOVA

MASTER IN COMPUTER SCIENCE AND ENGINEERING

NOVA University Lisbon
September, 2024

SLVideo: A Sign Language Video Moment Retrieval Framework

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.1.5) [1].

Para o meu avô, Domingos.

Acknowledgements

I am truly grateful to my adviser, Professor João Magalhães, who has guided me through this last year with his teaching and support to develop a valuable thesis that I believe in. I would also like to thank to the members of the Nova Search team, who welcomed me into their group and were always available to lend a helping hand.

My appreciation extends to the friends that I have made along my academic journey, who have accompanied me over the last five years, not only in classes and group projects but also in memorable social gatherings and moments of both triumph and challenge.

To my childhood friends, both from school and from my orchestra, who I have always been able to rely on, who have always supported me in my difficult times, and who have always been there to cheer on my successes.

Last, but in no way least, to my parents, my sister, my grandparents, my family, for their immeasurable support and love, for which there are no words I could say to express my gratitude, for your endless patience, understanding, and sacrifices, to which I owe everything.

”

“Learning never exhausts the mind.”

— Leonardo da Vinci

Abstract

Sign Language Recognition has been an increasingly studied and developed subject throughout the years to help deaf and hard-of-hearing individuals in their social interactions in everyday life. These technologies employ manual sign recognition algorithms; however, the majority of them lack the capacity to recognise facial expressions, which are also an essential part of sign language as they allow the speaker to add expressiveness to their dialogue or even change the meaning of certain manual signs. For Portuguese Sign Language Recognition software this is no exception. This dissertation introduces SLVideo, a video moment retrieval system for Sign Language videos that incorporates facial expressions, addressing the gap in existing technology by focusing on both hand and facial signs. The system extracts embedding representations for the hand and face signs from video frames to capture the language signs in their entirety. This enables users to search for a specific sign language video segment with text queries or to search by similar sign language videos. To evaluate this system, a collection of eight hours of annotated Portuguese Sign Language videos is used as the dataset, and a CLIP model is used to generate the embeddings. The initial results are promising in a zero-shot setting. Additionally, SLVideo allows users to edit existing annotations and create new ones, making it a collaborative tool for annotators working with the same videos.

Keywords: Sign Language Recognition, Facial expressions, Portuguese Sign Language, Video moment retrieval

Resumo

O Reconhecimento de Língua Gestual tem sido um tema cada vez mais estudado e desenvolvido ao longo dos anos para ajudar as pessoas surdas e com dificuldades auditivas nas suas interações sociais do dia-a-dia. Estas tecnologias recorrem a algoritmos de reconhecimento de sinais manuais; no entanto, na sua maioria, carece do reconhecimento de expressões faciais, que são também uma parte essencial da língua gestual, pois permitem ao falante acrescentar expressividade ao seu diálogo ou mesmo alterar o significado de determinados sinais manuais. No caso do software de reconhecimento de Língua Gestual Portuguesa isto não é exceção. Esta dissertação apresenta o SLVideo, um sistema de recuperação de momentos em vídeos de Língua Gestual Portuguesa que incorpora expressões faciais, colmatando a falha existente na tecnologia atual ao focar-se tanto nos sinais manuais como nos faciais. O sistema extrai representações em *embeddings* para os sinais manuais e faciais a partir dos *frames* de vídeo para captar os sinais linguísticos na sua totalidade. Isto permite aos utilizadores procurar um segmento específico de vídeo de língua gestual com consultas de texto ou procurar por vídeos de língua gestual semelhantes. Para avaliar este sistema, é utilizada uma coleção de oito horas de vídeos anotados de Língua Gestual Portuguesa como conjunto de dados, e é utilizado um modelo CLIP para gerar as *embeddings*. Os resultados iniciais são promissores num cenário de zero-shot. Além disso, o SLVideo permite aos utilizadores editar anotações existentes e criar novas, tornando-o numa ferramenta de colaboração para anotadores que trabalham com os mesmos vídeos.

Palavras-chave: Reconhecimento de Língua Gestual, Expressões faciais, Língua Gestual Portuguesa, Recuperação de momentos de vídeo

Contents

List of Figures

List of Tables

Acronyms

ANN approximate nearest neighbour *(pp. 52, 66)*
ASL American Sign Language *(pp. 10, 11)*

CLIP Contrastive Language-Image Pre-training *(pp. 11, 18, 36, 40, 49, 50, 57)*
CNN Convolutional Neural Networks *(pp. 9, 25, 26)*

EAF ELAN Annotation File *(pp. 30, 32, 33, 41)*

FER Facial Expression Recognition *(pp. 11, 12)*

k-NN k-nearest neighbors *(pp. 40, 52, 53, 66)*

Libras Brazilian Sign Language *(p. 8)*
LLMs Large Language Models *(pp. 20–22, 26, 27)*
LSA Latent Semantic Analysis *(p. 10)*
LSTM Long Short-Term Memory *(pp. 25, 26)*

METD multiple expression text descriptors *(p. 12)*
MIL multiple instance learning *(p. 17)*

NLP Natural Language Processing *(p. 21)*

OVIS ontology video surveillance indexing and retrieval system *(p. 15)*

PRVR Partially Relevant Video Retrieval *(p. 17)*
PSL Portuguese Sign Language *(pp. 3, 4, 7, 8, 29, 30, 32, 34–36, 39, 52, 55, 57)*

RNN Recurrent Neural Networks *(p. 25)*

SLR Sign Language Recognition *(p. 1)*

T2V Text-To-Sign-Video *(p. 19)*

V2T Sign-Video-To-Text *(p. 19)*
VCMR Video Corpus Moment Retrieval *(pp. 16, 17)*
VCR Visual Commonsense Reasoning *(pp. 22, 25)*
ViGA video moment retrieval via Glance Annotation *(p. 18)*
VMR Video Moment Retrieval *(pp. 1, 16–18, 27, 41)*
VQA Visual Question Answering *(pp. 22, 25)*

1

Introduction

Consider a world where communicating with other people is a constant challenge. This is the reality for the deaf and hard-of-hearing community. In their daily lives, deaf and hard-of-hearing people struggle to communicate with hearing people who do not know sign language, without existing sign language interpreters in most places. Writing is not always a solution, because the "writing language" is like a whole different language to deaf people, so learning it is not that straightforward.

The emergence of Sign Language Recognition (SLR) and Generation technologies helps to solve this problem, with some teams developing software that can translate speech to sign language by an avatar, or the opposite, translating a sign language video to text by using video recognition. These technologies are very helpful for deaf and hard-of-hearing people, easing their daily communication with hearing people and their usage of social services, like going to the hospital, a restaurant, etc. Besides its social advantages, this software would also simplify the education of non-hearing children, as it would translate what the teacher is saying into the appropriate sign language.

SLR can also be applied in visual question answering or Video Moment Retrieval (VMR) tasks with sign language videos (Figure 1.1). Visual question-answering is typically a task where the user gives a natural language query to the system regarding an image and/or video, while a VMR task is when a user also gives a natural language query regarding a video and the system must return the appropriate video segment that corresponds to the query.

Software capable of finding a specific video segment in sign language videos where the searched phrase or word is signed could be very useful for non-hearing and hearing people in several ways. For instance, a student learning sign language could search for a specific word or phrase in a video in order to see how its sign is performed, or a non-hearing person learning how to read could find the corresponding sign for an unfamiliar word or phrase. It also helps communication between hearing and non-hearing people, as a person who doesn't speak sign language could search for a phrase and show the corresponding video to a non-hearing person in order to communicate with them.

Given the complexity of this technology, effective SLR software has to support the

Figure 1.1: Example of a video moment retrieval task with sign language videos using a similarity score [3].

recognition of all the nuances of sign language, including non-manual signs such as facial expressions and head and shoulder movements. These elements are an essential aspect of sign language, due to their grammatical and expressive functions in the dialogue. Facial expressions, specifically, can distinguish what type of phrase is being said, add intensity, and even change the meaning of a manual sign [2].

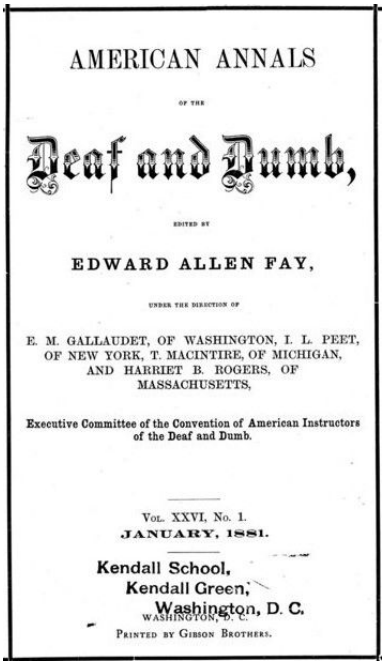
1.1 Historic Context

The importance of facial expressions in sign language comes from a difficult time in the sign language area, originated by the Milan Conference in 1880 [4]. This conference imposed that deaf people should be able to communicate without using hand signs, but by speaking like a hearing person does, replacing manual education (sign language) with oral education (oralism). This decision defended that in this way non-hearing people would be more easily included in society and it was accepted by some European countries and the United States, leading to deaf teachers losing their jobs and a general decrease in deaf professionals. At that time, there was a magazine called "American Annals of the Deaf and Dumb" (Figure 1.2) which published an article about this conference[1].

More recently, this was seen as an act of discrimination and disregard for human rights, as it made it more difficult for these people to communicate, resulting in consequences contrary to what was intended. As such, in 2010, a formal apology was made at the 21st International Congress on Education of the Deaf by the board.

The development and evolution of facial expressions in sign language was an inevitable outcome of the Milan Conference, now having an important role in the communication

[1] American Annals of the Deaf and Dumb Vol.26 No.1


AMERICAN ANNALS

OF THE

Deaf and Dumb,

EDITED BY

EDWARD ALLEN FAY,

UNDER THE DIRECTION OF

E. M. GALLAUDET, OF WASHINGTON, I. L. PEET, OF NEW YORK, T. MACINTIRE, OF MICHIGAN, AND HARRIET B. ROGERS, OF MASSACHUSETTS,

Executive Committee of the Convention of American Instructors of the Deaf and Dumb.

VOL. XXVI, NO. 1.
JANUARY, 1881.

WASHINGTON, D. C.
PRINTED BY GIBSON BROTHERS.

Figure 1.2: Cover of the American Annals of the Deaf and Dumb magazine where it was published an article about the Milan Conference.

of sign language users as it can add intensity to what is being said and/or change the meaning of a manual sign.

1.2 Motivation

Despite the importance of facial expressions in communication through sign language, most of the approaches for Sign Language Recognition focus on hand gestures and put facial expressions as a low priority research, which is very concerning as in that way it will lose a lot of linguistic characteristics of the signs performed, not being as helpful as it should.

For a video moment retrieval task with sign language videos, the lack of facial expression recognition could lead to a wrong video segment being returned for a query where the facial expression has an important role.

This is a very complex and interesting topic, as it faces several challenges, like the diversity and complexity of the existing sign languages, the lack of large and annotated datasets that include annotations for the facial expressions, and the need for multimodal and cross-modal understanding for interpreting the videos and the text query.

1.3 Objectives

This dissertation is part of a project that aims to develop a Sign Language Recognition system for videos where the main element is Portuguese Sign Language (PSL), with an

emphasis on the recognition of facial expressions, in order to address the current lack of software for this purpose.

We will be focusing on searching methods for PSL videos, exploring large language models and how they can be used to locate the relevant segments of videos when the query can be expressed in text or video. For this, we will need to explore video processing techniques, more specifically video indexing, video structure, and video moment retrieval. We'll also need to have a basic knowledge of sign language and the role of facial expressions.

As previously stated, the main element of the PSL videos to be explored are the signs where the facial expression is essential to understand the sign being performed, whether it is an addition to a manual sign or it is the sign itself, i.e., the sign is composed only with facial expressions with no manual signs. Additionally, we will address the shortage of annotations regarding PSL videos and the nuances of facial expressions in sign language.

1.4 Outcomes and Contributions

With the previously defined objectives in mind, we developed SLVideo, a video moment retrieval system for Sign Language videos, with the following features:

- The user can **search for a sign using a text query** and get a set of the **corresponding video segments** to that query;
- There are **annotated Portuguese Sign Language videos** available for the users to watch and explore their annotated facial expression signs;
- The user can **edit the annotations and create new ones**;
- A **thesaurus** feature that allows the users to **search for similarly performed signs** from the ones that are retrieved when performing a query.

A ready-to-use web application was developed and is accessible through the URL slvideo.novasearch.org, with a repository available in GitHub, that allows users to explore all the SLVideo features. We also developed a paper to be submitted regarding SLVideo available in arxiv. All of these resources can also be accessed through the SLVideo GitHub page.

1.5 Document Structure

This dissertation adopts the following structure:

- **Chapter 1, Introduction**: This section introduces the dissertation, explaining its background, motivation, and objectives, giving a general nuance of what was developed.

- **Chapter 2, Related Work**: This section explores all the areas of work that are relevant to this dissertation, including a basic explanation of each one and its state-of-the-art. The areas explored are sign language linguistics and the role of facial expressions in it, video processing, language models and their compatibility with videos, and sign language recognition and generation.
- **Chapter 3, SLVideo: System Architecture and Implementation**: This section explains how SLVideo was developed, exploring its architecture and implementation, all its components, processes and algorithms, the dataset of sign language videos used and the challenges faced throughout this development.
- **Chapter 4, Evaluation**: This section explores the methodology used to evaluate the developed system, the results of that evaluation, the interpretation of those results and how they contributed to the development of the version of SLVideo presented in this dissertation.
- **Chapter 5, Conclusions**: This final section concludes this dissertation, highlighting the most significant results and insights gained from it, the impact of this work in the world and sign language recognition area, the limitations and challenges faced through its development, and future work that can be done to expand SLVideo.

2

Background and Related Work

This chapter will present areas of work that are related to this dissertation, explaining its bases and exploring the existing work. First, we will provide a brief overview of sign language linguistics and the importance of facial expressions in it. If the reader wants to know more about these two topics, mainly for Portuguese Sign Language, we recommend reading the papers [5–7]. Then, we will explore existing work in computer science relevant to this thesis, specifically, we will focus on video processing, including video structure, video indexing, and video moment retrieval, and also examine the video and language models area, including large language models, cross-modal encoders, and video captioning. Finally, we will delve into the areas of sign language recognition and generation.

2.1 Sign Language Linguistics

Sign language is a combination of manual movements, body movements and facial expressions used by non-hearing people to express themselves, having its own vocabulary and grammatical rules. It is not a direct translation of the spoken language, such that the written form of the spoken language does not correspond to the written form of the sign language, and countries that speak the same language might not have the same sign language (for example, the American Sign Language is very different from the British Sign Language, not being mutually understandable, despite both countries speaking the same language). Some articles explore and explain the linguistics of sign language, like [2, 8–10].

Words are expressed by a combination of manual and non-manual signs, such that both have parameters that define what is being expressed. For manual gestures, the form that each hand assumes, the orientation of the palm of the hand, the location where the gesture is being made, and the movement of the hands (its speed and path) define the gesture. Additionally, there is a distinction between a dominant and a non-dominant hand for each sign language user. For the non-manual gestures, there is body movement (mainly head and shoulder movements) and there are facial expressions (such as eyebrow

motion and lip-mouth movements) that add a sense of emotion to the gesture. For words that don't have a translation for sign language, a technique called fingerspelling is used, consisting of the gesturing of the letters that compose the intended word in the local spoken language.

To represent the meaning of the signs in a written form, gloss annotation [8] can be used, although this does not provide any information about the execution of the sign, only its meaning. Despite being an approximation of a solution for writing sign language, glosses only embody the manual aspect of the sign, so it may not represent its full meaning by not including the non-manual signs, leading to a loss of information.

2.1.1 Portuguese Sign Language

The education of the deaf in PSL appeared in the 19th century when the Swedish professor Per Aron Borg came to Portugal at the request of King João VI and founded a school for deaf children to learn a communication system. PSL also suffered from the consequences of the Milan Conference of 1880, but it was rescued by the University of Lisbon and the Portuguese Association of the Deaf at the end of the 20th century. The most renowned dictionary for PSL is the 'Dicionário de Língua Gestual Portuguesa' [11], which is complemented by some existing manuals [7]. The papers [5, 9, 10], also have some explanation about the linguistics and history of the Portuguese Sign Language.

Most nouns have an associated sign to express it. Only animated beings have a gender, and when one wants to specify the gender it has to do an extra sign as a prefix representing 'female' or 'male', and if no prefix is added it is assumed to be male, except for some cases where the default noun is female (like in 'enfermeira', which translates to 'nurse'). To represent plural cases, the sign can be repeated (repetition), can be made with both hands (reduplication) or can specify the quantity after the noun (incorporation), being that each of the techniques depends on the noun in question. For forenames, fingerspelling is used.

The verbs are usually said in their infinitive form. If the tense is relevant, the usage of an imaginary temporal line is necessary, being expressed by non-manual signs using the eyes, eyebrow and upper body movement (for example, if the verb is representing an action happening at that moment, the sign should be performed right in the front of the signer). Another way is to add a time adverb to the sentence, also according to the temporal line. A verb can also express the manner, duration and repetition of the action by modulating the movement of the sign (distinguishing "walking" and "stumping", for example). The context also affects how the verb is signed (for example, 'to eat' can be signed differently depending on what is being eaten).

2.2 Facial Expressions in Sign Language

In sign language, facial expressions are part of the non-manual signs, being extremely important since they have grammatical, lexical and affective roles in it, giving more

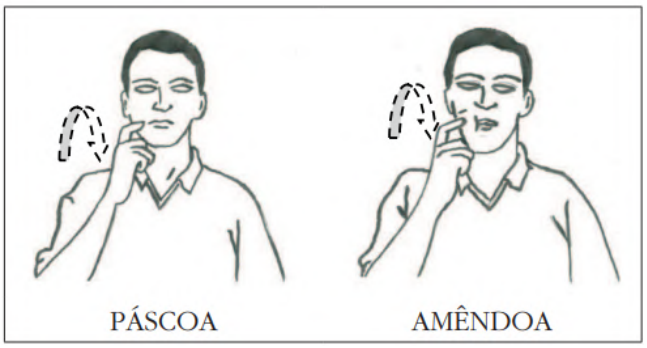


Figure 2.1: Despite having the same manual gesture, the facial expression changes the meaning of the whole sign from 'Páscoa' (Easter) to 'Amêndoa'(Almond) [6].

expressiveness and clarity to the sign that is being executed. Some articles, like [2, 6, 8], explore and explain the linguistics of facial expressions in sign language.

For the grammatical roles, facial expressions can distinguish the type of phrase that is being said, i.e., if it is affirmative, imperative, exclamatory, conditional or interrogative (differentiating a WH-question of a Yes/No question), and can also express quantity and temporal relation. There are manual signs that are executed the same way but the whole meaning changes depending on the facial expression, like for the words 'Páscoa' ('Easter') and 'Amêndoa' ('Almond') in Portuguese Sign Language (Figure 2.1). For the affective roles, they can give emotion to what is being said, for example, there is only one main sign for "large" but depending on the facial expressions, mouthing and the size of the sign it can lead to different levels of largeness (big, large, huge, etc).

2.2.1 Facial Expressions in Portuguese Sign Language

In PSL, facial expressions are very important in its linguistics, as they also have the same roles as in the generality of the existent sign languages, with some few differences [6]. Besides the general grammatical roles of facial expressions in sign languages, in PSL they can also make important phonological distinctions in signs that only have a minimal change that differentiates them, and can also define the grammatical forms of adverbial, giving a sense of state ("slower", "faster", "delicately", etc), and adjectival modifiers, giving distinctions of quality and quantity.

There are several non-manual sign movements in PSL, as shown in Table 2.1, that, besides the facial expressions, also include head and body movements.

2.2.2 Facial Expressions in Brazilian Sign Language

Silva and Costa [2] focused on the development of an automatic sign language recognition system with special attention to facial expressions, using Brazilian Sign Language (Libras). However, they found that there still isn't much documentation on the list of facial expressions in Libras and how to combine them with manual movements, so they proposed a list of non-manual signs and their classification, which can be seen in Table 2.2.

Face	Head	Face and Head	Torso
Frowning (closed) and raised (open) eyebrows; Eyes;	Forward and downward movement (yes) and sideways (no); Forward, sideways and backward tilt;	Head projected forward; Eyes slightly closed and eyebrows frowning or raised; Head projected backwards and eyes; Head rotating;	Alternating shoulder balances; Simultaneous shoulder balances; Single shoulder balances;

Table 2.1: Non-manual sign movements in Portuguese Sign Language [6].

There are also affective facial expressions and grammatical facial expressions. The authors figured that it is possible to categorize the latter in grammatical facial expressions for sentence (define the type of sentence being said), grammatical facial expressions of intensity (adds a quantifier to the sentence) and grammatical facial expressions of distinction (distinguishes signs with the same manual gestures).

2.3 Sign Language Recognition

Automatic Sign Language Recognition helps the communication between hearing people and deaf or hard-of-hearing people, translating sign language into written or spoken sentences of the desired language. This technology requires special attention to the hands and finger gestures, which are harder to detect and also transmit a lot of information. The more samples to be analysed the better, because no sign language speakers speak the same way, changing the speed, their dominant side, body shape and other factors. However, getting that many samples is expensive.

One possible approach to sign language recognition is to adapt skeleton-based action recognition to this area, which consists of the recognition of the human body motions using skeletal joint data captured by motion capture systems. Jiang et al. [12] suggests a multi-modal approach to skeleton-based sign language recognition, using a spatio-temporal graph that connects human body keypoints extracted from a video, including facial landmarks, hands, feet and body keypoints. To model the dynamics of this graph, it is proposed a Sign Language Graph Convolution Network, which generates bone and joint motion data. Besides that, a Separable Spatial-Temporal Convolution Network is proposed to recognize sign language from whole body features, extracting them from the chosen keypoints and subjecting them to a series of convolution methods. Three-dimensional Convolutional Neural Networks (CNN) are employed for the other modalities of RGB frames, optical flows, depth HHA and depth flows in order to boost performance. All the modalities used are ensembled to boost the accuracy of the system and the results obtained are very satisfactory.

AFE	Left / Right eyebrow raised; Raised eyebrows and wide-open eyes; Raised eyebrows, wide-open eyes and open mouth; Slightly closed eyes and crooked mouth up; Smile with apparent teeth; Smile with apparent teeth and open mouth; Lowered eyebrows and crooked mouth down; Frown and contraction of the upper lip; Crooked mouth up laterally.		
GFE	GES	WH	Brief and upward movement of the head and frown.
		YN	Brief and upward movement of the head and raised eyebrows.
		DQ	Frown, slightly closed eyes and contracted lips.
		T	Brief upward and forward movement of the head, raised eyebrows, open mouth, projected lips; Quick nod, brief upward movement and wide open eyes; Quick nod, brief upward movement, raised eyebrows and wide open eyes; Quick nod, brief upward movement, raised eyebrows, open mouth and projected lips.
		N	Crooked mouth down; Quick nod, frown and crooked mouth down; Head balancing sideways.
		A	Balance back and forth of the head.
		CC	Brief and upward movement of the head and raised eyebrows.
		F	Brief upward and forward movement of the head, raised eyebrows, open mouth, projected lips; Quick nod, brief upward movement and wide open eyes; Quick nod, brief upward movement, raised eyebrows and wide open eyes; Quick nod, brief upward movement, raised eyebrows, open mouth and projected lips.
		RC	Raised eyebrows.
	GEI	Frown < Frown and Slightly closed eyes; Inflated cheeks and semi-open mouth < inflated cheeks, semi-open mouth and frown; Contracted cheeks and frown < contracted cheeks; Contracted lips and frown < contracted lips; Projected lips and frown < projected lips; Open mouth and frown < open mouth; Crooked mouth up< Smile with apparent teeth; Quick nod < balance back and forth of the head;	
	GED	Left eye closed; Inflated cheeks; Only right cheek inflated;	

Semantic Functions: AFS- Affective Facial Expression; GFS- Grammatical Facial Expression; WH- WH-question; YN-Yes/No questions, DQ- Doubt question, T- Topic, N- Negation, A- Assertion, CC- Conditional Clause, F- Focus, RC- Relative Clause.

Table 2.2: Libras non-manual signs classification defined by Silva and Costa [2].

There are some dictionaries[1] available to translate sign language to the respective spoken language, although with limitations. Some accept textual queries and return videos or images, but that requires the user to make a text description of the sign or its meaning, which is not intuitive. Others allow for a video query, but there are no guarantees that a user will replicate the sign accurately or that the system will recognize it. Another interesting approach to sign language recognition is the use of feature-based search in dictionaries, like the one Bragg, Rector, and Ladner [13] propose, which is a dictionary for American Sign Language (ASL) called ASL-Search, an ASL-to-English dictionary that allows users to search for a sign by selecting a set of features that describe it, such as hand shape, orientation, location and movement. The system stores queries from previous users in a matrix of feature frequencies for each sign and uses Latent Semantic Analysis (LSA) to learn from those queries and improve the search results by reducing the noise

[1] Online Sign Language Dictionary

in the data and the number of features needed when comparing a query to the database. The authors also developed ASL-Flash, an educational tool that presents a series of ASL flashcards and simultaneously gathers additional query data for ASL-Search. This way, it demonstrated the performance of ASL-Search in a proof of concept for the system design, which results proved that ASL-Search surpasses comparable existing dictionaries and will continue to improve over time with usage.

2.3.1 Facial Expressions Recognition

If sign language recognition systems could also be capable of recognizing facial expressions it would be able to do more precise and expressive translations of the signs performed, as the lack of facial expressions could lead to mistakes.

To recognize facial expressions in sign language, it is necessary to take advantage of facial landmarks and facial action units [8]. Facial landmarks are points on the face that encode the contours of key facial features such as the eyes, nose and mouth. Facial action units describe the intensity of facial muscle movements and are used in psychology and affective computing to understand emotions expressed through facial expressions. As such, both of these features are essential to these tasks as they capture the nuances of facial expressions in a very reliable way.

2.4 Visual Descriptors - CLIP

Radford et al. [14] studied the behaviours of image classifiers trained with natural language supervision, a method that can be used for training machine learning models to predict which caption goes with which image. For this, the authors created a dataset of 400 million (image, text) pairs, collected from the internet, and created a method of learning visual models from natural language supervision called Contrastive Language-Image Pre-training (CLIP), which learns to predict if an image and a text snippet are paired together in its dataset by maximizing the cosine similarity for features of the matched image-text pairs while minimizing that of all other negative pairs. This model is also used to perform zero-shot classification, by using the names of all the classes in the dataset as the set of potential text pairings and predict the most probable (image, text) pair. After evaluation, the authors find that CLIP can perform zero-shot transfer much better than the image caption baseline. This method can be applied not only to images but also to video, by substituting the (image, text) pairs for (frame, text) pairs.

2.4.1 CLIP for Facial Expressions

There has also been some work on CLIP applied to facial expressions. Li et al. [15] proposed CLIPER, a framework for both static and dynamic Facial Expression Recognition (FER) based on CLIP. While CLIP can categorize clear and uniform textual definitions (e.g., "dog" and "boat"), that is not the case for facial expression categories, such as "happy"

or "angry", whose textual definitions are more abstract and compound. To tackle this issue, CLIPER can automatically learn a set of expression text descriptors for each facial expression without adding any expression-related textual prior. This model has two training phases: first it learns the proposed multiple expression text descriptors (METD) for each expression class, and, in the second stage, the image encoder is fine-tuned to extract more discriminative facial expression features. The usage of METD helps CLIPER to learn fine-grained expression representations and be more interpretable than other FER approaches. After experiments on FER benchmarks, the effectiveness of CLIPER is verified by achieving state-of-the-art performance, being able to serve as a strong baseline for this type of task.

2.5 Video Processing

The quick and drastic advances in artificial intelligence help the development of multimedia computing, more specifically video and image analysis. This type of analysis requires the extraction of multimedia metadata, such as colour, shapes and texture, and for videos, it also requires the detection of temporal changes throughout the frames and the understanding of audio.

As this dissertation will require the analysis and interpretation of several sign language videos, it is important to understand how to process those videos.

2.5.1 Video Structure

A raw video is an unstructured data stream composed of a sequence of shots, which can be decomposed into keyframes. These shots can be grouped by similarity, and semantically related shots can be merged to form scenes (Figure 2.2). Besides these concepts, story (a series of shots that capture a continuous action) and subshots (a portion of a shot characterized by a single camera motion) are also part of the structure of a video.

Most of the current systems rely on low-level primitives (like colour, texture, and motion) to characterize a video, but, despite its efficiency, those do not include the video semantics, so the system has to make the bridge between the low-level primitives and the high-level semantics [16]. Having a video with a more detailed structure is very helpful to its understating, increasing the context analysis performance and facilitating feature extraction.

While shots are marked by physical boundaries, scenes are marked by semantic boundaries. Effective shot boundary detection, scene grouping and key frame extraction mechanisms are essential to good video analysis. Shot boundary detection is the process of identifying the boundaries between two consecutive shots, allowing a sequence of frames to be organized into a collection of shots. The transition between consecutive shots in a video can be classified into two main types: cut and gradual transitions. A cut transition signifies an abrupt change between two shots, while a gradual transition implies

a slow, special effect-based change between the two shots [17]. Automatic shot boundary detection can be divided into five categories: pixel-based, statistics-based, transform-based, feature-based and histogram based. The histogram-based approach is recognized as achieving a good trade-off between accuracy and speed by some researchers [18].

A keyframe is the frame of a shot that best represents its content. Several keyframes can serve as reference points for video data, enabling operations like indexing and browsing. After defining the shot boundaries, keyframes can be extracted. Some more simple approaches just extract the first and last frames of each shot, although there are two more complex main approaches: analysis-based and clustering-based [17]. Analysis-based consists of extracting keyframes by evaluating aspects such as the quality and attractiveness of frames, and in clustering-based a clustering process is performed and the cluster centroids are designated as the keyframes. More recently, keyframe extraction has been formulated as a learning task, such that the model is trained to recognize the frame representativeness using image quality, user attention, and visual details as features.

Scene grouping is the organization and grouping of shots based on the results of the shot boundary detection, which is extremely important in the understanding of a video. Shots from the same scene might not be visually similar but have to be grouped since they are semantically related. For example, usually in a movie when there is a conversation between two persons the shots will switch back and forth between them, and, although there exists two separate groups of shots, one for each person, these are semantically related, being part of the same scene. As such, scene grouping considers the content similarity and the temporal continuity of the shots when executed. The content similarity is based on the visual, audio and text features of the video, such that shots from the same scene should have similar contents, whereas the temporal continuity indicates that shots from the same scene are close to each other temporally [17]. As mentioned in [19], there is a technique called silence detection that uses the audio characteristics to check if there is a significant decrease in the audio volume to detect possible scene changes, which is not a very reliable approach but if combined with other techniques can possibly increase their performance.

Rui, Huang, and Mehrotra [18] explained their approach to scene grouping. The authors use the first and last frames of every shot as the keyframes. The temporal information of the shot is characterized by the extracted shot activity measure and the spatial information by the extracted visual features of the keyframes. Then they group the similar shots, since these are the most likely to be in the same scene, using a time adaptive grouping approach based on the content similarity and the time continuity. Finally, semantic-related shots are grouped into a scene using a scene structure construction method. This simple approach was proven to be effective and capable of creating accurate scenes.

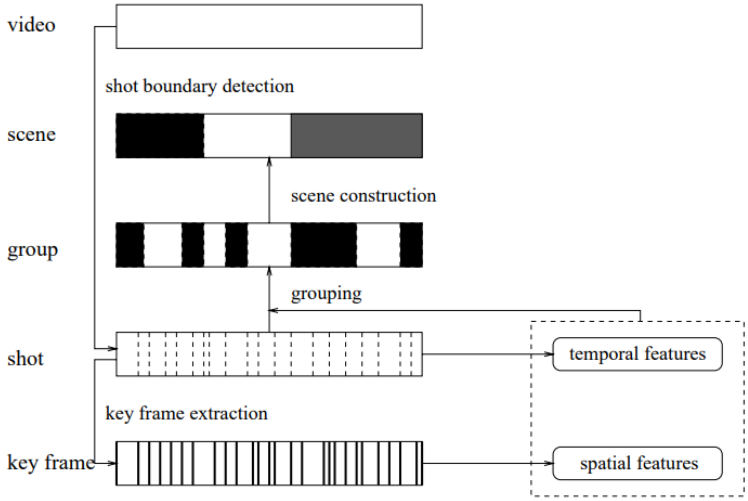


Figure 2.2: Rui, Huang, and Mehrotra [18] hierarchical video representation.

2.5.2 Video Indexing

With the exponential increase of recorded data in size and numbers in the last few years, the need to search for this kind of data and retrieve information from it has made multimedia indexing a major research topic.

Video indexing consists of analysing a video and creating an index of its content to make it more easily searchable and discoverable, involving the parsing of the video data, identification of keyframes, recognition of text, objects and audio, etc. So, in the case of a video, an index is like a catalogue with the details of the content of each video that helps users find what they're looking for more efficiently. Search engines like Google use this process to provide the content that a query written by the user is asking for.

For video indexing based on visual content, there are two main approaches: image classification and image description [20]. Image classification consists of assigning predefined tags to each frame, keyframe or scene of a video, with some improvements including salient objects detection (tagging the bounding boxes in video frames) and image segmentation (tagging free-form image regions), being that, regardless of the method used, the resulting index includes a collection of time codes and categories for each segment of the video. Image description involves generating natural language text annotations based on video frame content, which is more suitable for general search engines but is not efficient for searching by examples (as required by MPEG-7 standard). These two approaches can also be combined to create a more robust and complete index.

Podlesnaya and Podlesnyy [20] suggests a graph structure for video indexing, where it represents a video as a graph with shots and tags as nodes, and categories, places, faces, salient objects, and spatial relations as edges, using the Neo4j graph-oriented database to store and query the video index, allowing querying the video index by keywords, synonyms, hypernyms, and spatial relations. After some testing, the results were satisfactory, as it efficiently retrieves information using complicated spatial and

temporal search queries and allows querying with the Cypher query language which can express logical and linguistic relations between video elements.

The major problem of video indexing, even if the indexing mechanism combined image classification and image description, is the semantic gap between the extracted low-level features and the information perceived by humans that are annotated. Several works suggest an ontology-based approach to overcome this issue, but building an ontology that covers a large set of events and objects is not an easy task.

Kazi Tani et al. [21] propose an ontology for the video surveillance domain, dividing it into four categories, that are linked together, for better coverage of all the objects and events that can happen in this kind of video: video actions, video events, video objects and video sequences. All the video segments are indexed with one or more concepts of the video events category, which is related to the video actions since an event is a composition of one or more actions. Video actions and video objects are also related, and so are video sequences and video objects. This ontology is composed also of the DataProperty, which concerns all the properties related to one or more concepts, and the ObjectProperty, which includes the concepts of the ontology interactions. This approach supports the event indexing process in the video surveillance domain, being the core module of the ontology video surveillance indexing and retrieval system (OVIS) proposed by the authors. The indexing process starts with the extraction and organization of the different blobs bounding boxes from the video sequence to create DataProperty and ObjectProperty and then indexes the video sequence into the fitting video event class according to its objects' behaviour with start and end event frames. Then, the video surveillance stream will be indexed and stored in the database. This way, when one desires to retrieve a video clip, one can just search for an event keyword (such as "walking", "running", and so forth). After testing, this system was proved to be very efficient and extendable.

Most of the video indexing systems are based on low-level features, neglecting the hypothesis that a user may want to search for a video based on what it makes them feel. For this reason, affective video analysis [22] (associating emotions to the video) has been more explored, including the viewer's point of view in the videos' indexes, which improves the user experience on video retrieval. There are two main approaches for this purpose: the usage of low-level features to find objective emotions, i.e., emotions that are expressed by the video itself, like dark colours describing a "sad" moment, and the usage of viewers' feelings to find subjective emotions. Objective emotions' drawback is that they might express the director's point of view and not the viewer's. Subjective emotions are nowadays difficult to collect, requiring access to the user's physical signals and facial expressions to deduce what they are feeling.

Furini [22] proposes a video indexing mechanism that combines objective and subjective emotions called ViMood. This system performs a low-level video/audio analysis to retrieve its objective emotions and uses the viewer's point of view annotations for the subjective emotions by giving them an interactive interface with several emotions where the user might press on one of them to express their feelings while watching that

specific video frame. ViMood allows users to browse for video material using general information (like genre, cast, director, etc.) but also using eight emotions ("joy", “trust”, “fear”, “surprise”, “sadness”, “disgust”, “anger” and “anticipation”), by creating an index that associates an emotion to every video segment that composes the main video.

So, as observed, there are several approaches to video indexing, being that a lot of the solutions opt to index the audio, the video, or the text modality exclusively, which can be limiting, although there are already some solutions opting for a multi-modal approach that merges all these modalities, giving more complementary and redundant information about the video content.

Snoek and Worring [19] presented an overview of the state-of-the-art multi-modal video indexing, proposing also a framework that does the indexing based on the video’s author perspective, distinguishing the intended semantics meaning, the content elements and the video structure. The authors mentioned that there are several challenges in multi-modal approaches, such as their temporal synchronization, which depends on which modality is the "main one" (for example, if audio is the main modality, the image frames are converted to milliseconds [23], but if the image is the main one, audio samples are assigned to image frames [24]). The multi-modal integration can be achieved in several ways, most of them being symmetric and non-iterated some following a knowledge-base approach and others a statistical approach, which is the most adopted ones as the usage of Hidden Markov Models and Bayesian network framework revealed to be very fulfilling. In the end, the authors state that it seems that the usage of advanced integration methods correlated with multi-modal video indexing is a promising approach to getting positive results in this area.

2.5.3 Video Moment Retrieval

VMR consists of getting the desired video segments given a natural language query, being a matter highly developed in machine learning and multimedia computing. Most of the methods for VMR use annotations including the video, the query, and the start and end of the interesting part of the video. With so many sign language videos to analyse, this kind of technology is essential in this dissertation, so exploring developed methods and related work will be very useful.

VMR is facilitated when done in trimmed and short videos, but there are also developments in retrieving specific moments from longer untrimmed videos and also from even a collection of videos, known as Video Corpus Moment Retrieval (VCMR), which requires a more complex approach. The generation of temporal labels [25] also facilitates this task, as the interesting parts of the video are already annotated when the query is given.

Zhang et al. [26] tries to improve VCMR by developing a Retrieval and Localization Network with the support of contrastive learning (ReLoCLNet). To learn the matching between query and video, there are two main approaches (Figure 2.3), one being unimodal encoding, where the text and the video are encoded separately and later executing feature

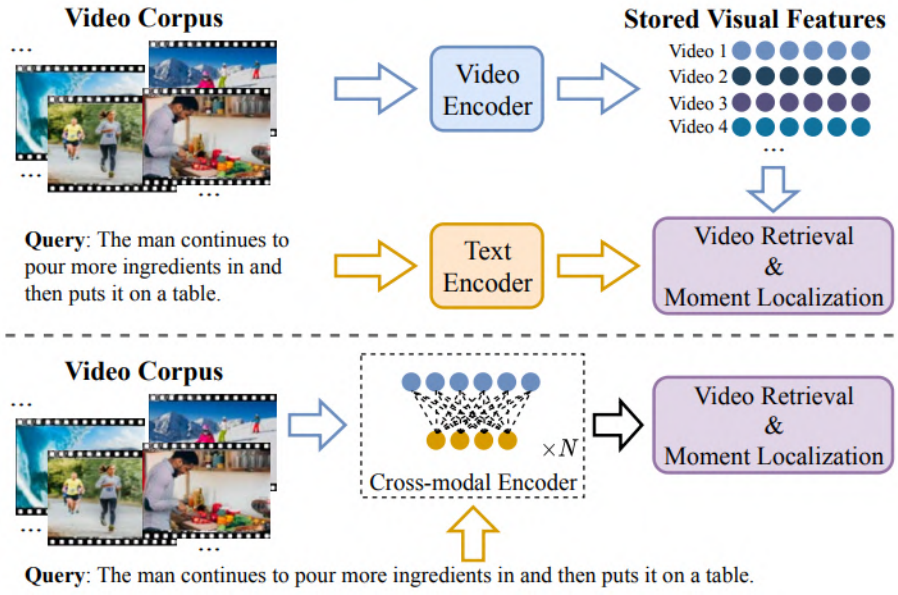


Figure 2.3: Two approaches to video moment retrieval using a corpus of videos: unimodal encoding vs cross-modal interaction learning [26].

fusing to learn the matching, and the other is cross-modal interaction learning, where a video is seen as a sequence of visual features and the query as a sequence of word features to learn their interactions, being the latter more accurate but also more expensive. The ReLoCLNet model uses unimodal encoding, developing a video encoder and a text encoder for effective feature encoding. It simulates cross-modal encoding by using contrastive learning in the training phase, as both goals are to match the video segments and the query. Once trained, the system is able to align the encoded video features and the text features. Finally, video retrieval and moment localization can be achieved by fusing the features.

Despite being simpler, having short-trimmed videos might not be the best approach, as a single video may not contain all the information that the query is trying to retrieve. Dong et al. [27] proposed the Partially Relevant Video Retrieval (PRVR) method, where it retrieves partially relevant videos, from a collection of untrimmed videos, that have a moment relevant to the given query (Figure 2.4), without any start/end timestamps of the moment. The developers formulate PRVR as a multiple instance learning (MIL) problem "where a video is simultaneously viewed as a bag of video clips and a bag of video frames". This method is a great addition to VCMR, such that it can filter the partially relevant videos, easing the retrieval of the specific moments of those videos.

Diwan, Peng, and Mooney [28] are implementing a VMR model using zero-shot learning, a machine learning concept where a model can understand and make decisions about data that has never been trained on. It starts by splitting a video into the different shot segments and, given a query, giving a similarity score to each segment. To obtain this similarity score, the developers explore two different methods, one being VideoCaptioning, where for each segment is produced a natural language caption and compared to the query,

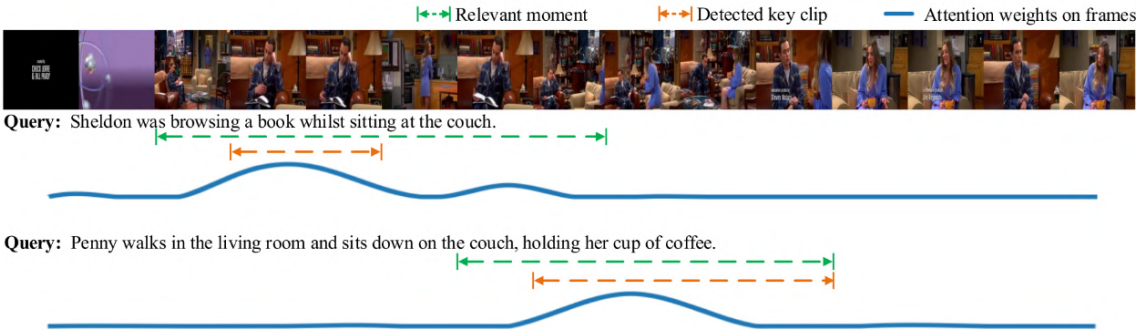


Figure 2.4: Video moment retrieval in an untrimmed video, where only a small part of the video is relevant to the query [27].

and the other being CLIP, where each frame is sampled and CLIP is used to embed the frames and the query, and the cosine similarity between the embedded frames and query is calculated. After computing the video segments and corresponding similarity scores, it is applied a post-processing step called SimpleWatershed, a simplified version of the Watershed algorithm [29]. The results obtained show that the best method is the zero-shot ShotDetect+CLIP+SimpleWatershed having an increase of 2.5x on all metrics.

2.5.3.1 Annotations in VMR

In deep learning, it can be developed a fully supervised VMR or a weakly supervised VMR [30], being the differences in the absence of start and end timestamps annotation on the elements of the dataset of a weakly supervised VMR, which is not influenced by the annotator's subjectiveness but can lead to misinterpretation of the video segment. Contrastive Learning can also be used to develop a VMR system, consisting of the comparison between similar and distinct samples, wanting to minimize the distance between similar samples and maximizing between distinct ones, i.e., minimizing the distance between related video frames and queries.

Cui et al. [30] propose glance annotation as a new annotation method, consisting of having a single timestamp between the start and end timestamps, lying in between fully and weakly supervised VMR, defending that this way one doesn't have to watch the whole video to find the interesting part, being also more informative than weak supervised VMR. This video moment retrieval via Glance Annotation (ViGA) method uses contrastive learning to train the model and Gaussian distribution to assign weights to the video clips and outperforms the existing weak supervised VMR approaches.

It is also possible to do VMR systems without having any annotations in the video, like using pre-trained models and object detectors to generate those annotations in the video and then match them with the query. Wang et al. [31] combined prompt learning, a machine learning technique where the user gives tips to the model to guide its predictions, with zero-shot learning for this purpose, using two main modules, the Proposal Prompt and the Verb Prompt (Figure 2.5). The Proposal Prompt is responsible for detecting actions

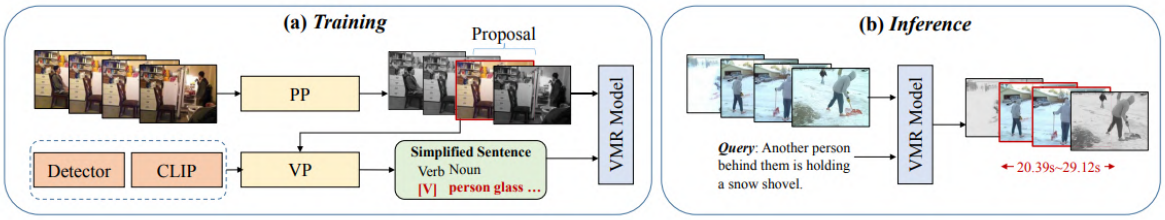


Figure 2.5: Overview of Prompt-based Zero-shot Video Moment Retrieval framework [31]. The Proposal Prompt (PP) detects events in the video, and the Video Prompt (VP) generates the pseudo-query to train the model. Then, the user-given query can be matched with the video moments.

or events in the video, by masking some frames with blank images, comparing each other and if there are significant changes then there probably is an interesting event in that interval. The Verb Prompt checks the detected actions and generates simple textual phrases describing those actions, and the moment localisation model is trained with those phrases. After the training phase, the query is easily matched with the video moments, because the model already knows how to combine visual features with different parts of a query.

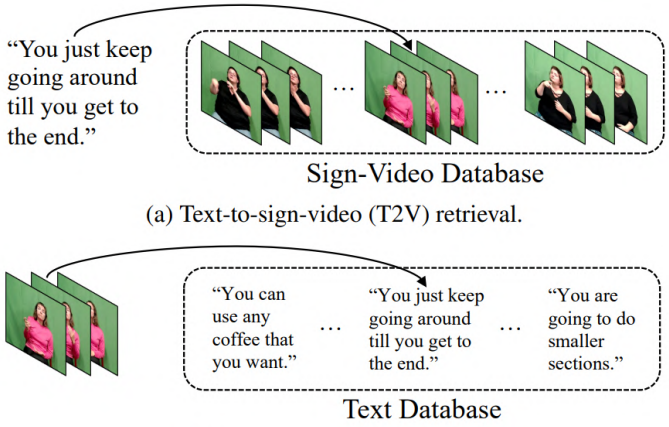
2.5.4 Sign Language Video Moment Retrieval

Sign language video moment retrieval can be composed of two sub-tasks (Figure 2.6): Text-To-Sign-Video (T2V), which finds the sign video with the sign content that best matches the written query, and the reverse task Sign-Video-To-Text (V2T), which finds the most relevant textual description given a query with a sign video.

This task has additional challenges in comparison to traditional video-text retrieval tasks, like the translation mappings between sign languages and spoken languages being very complex, as their modalities and grammatical structures are not the same, the sign language datasets available are much lesser and smaller and some even lack sign language data annotations, and it's harder to get a good sign embedding than a normal action embedding due to the necessity for the models to distinguish fine-grained gestures and actions.

Duarte et al. [3] address the task of sign language video moment retrieval with free-form textual queries, using the How2Sign dataset of American Sign Language [32], by learning a pair of encoders which map each signing video and text into a common real-valued embedding space.

The authors propose the SPOT-ALIGN framework to tackle the lack of sign annotations in the used dataset, which provides videos with only the corresponding written English translations, and increases the quality of the video embeddings. To do this, sign spotting techniques based on mouthing cues [33], which uses the mouth movements of the signer to identify signs that correspond to words in the written How2Sign translations, and



(a) Text-to-sign-video (T2V) retrieval.

(b) Sign-video-to-text (V2T) retrieval.

Figure 2.6: Illustration of (a) Text-To-Sign-Video retrieval and (b) Sign-Video-To-Text retrieval [35].

dictionary examples [34], which uses a collection of video examples of individual signs as visual queries to search for signs in the continuous test videos, are used to obtain candidate annotations, and then the video embeddings are retrained and the dictionary examples are re-queried to increase the annotation yield. It also used a text-based retrieval approach to complement the cross-modal retrieval approach and boost the overall performance. The experiments executed demonstrated the value of this framework in producing effective sign video embeddings for retrieval.

Considering the linguistic characteristics of sign language, Cheng et al. [35] approaches sign language retrieval both as a video-text retrieval problem and a cross-lingual (sign-to-word) retrieval problem and proposes a framework called domain-aware sign language retrieval via Cross-lingual Contrastive learning (CiCo).

CiCo is divided into two disjoint parts. First, the sign language videos are served as input to a sign encoder that pre-extracts their vision features. This encoder is obtained through transfer learning, by adapting a domain-agnostic sign encoder[36], pre-trained on large-scale sign videos, to the target domain and fine-tuning a domain-aware sign encoder on pseudo-labeled data from target datasets. Then, the extracted vision features and their corresponding texts are encoded in a joint embedding space using a cross-lingual contrastive learning module, allowing for the identification of fine-grained cross-lingual mappings. This approach outperforms the SPOT-ALIGN method by large margins on the How2Sign dataset.

2.6 Video and Language Models

With the rising development of Large Language Models (LLMs), the integration of video into this technology has also been highly explored as it opens opportunities for a lot of useful functionalities. For this, it is necessary to understand LLMs and how to combine

both text and video modalities.

The developments in LLMs and their capability to understand text, images and videos are also very important for this dissertation, as it will be necessary to interpret the sign language videos and produce text based on them.

In the case of this dissertation, the language can be either spoken language or sign language, although this section will be mainly focused on spoken language, as sign language interpretation is explored in section 2.3.

2.6.1 Large Language Models

LLMs are a consequence of the exponential advances in Artificial Intelligence technologies, in particular in the Natural Language Processing (NLP) area. These are deep learning models that are trained on a massive amount of data, achieving a great language understanding and generation, being able to generate human-like text, answer questions and complete tasks requested by the user on any topic with high accuracy, like summarizing a document, solving mathematical problems, generating a cooking recipe, and so much more. To maintain a conversation with the users, a Large Language Model has also to be capable of keeping track of the dialogue history, as some phrases can be misunderstood without the context (like the question "What is it used for?", where the AI does not know what 'it' is without the rest of the dialogue). For their capabilities, LLMs have gained a lot of popularity in the last few years with the emergence of user-available systems like OpenAI's ChatGPT and Microsoft's Copilot. These systems can be applied in many fields, including healthcare, finance, entertainment, customer support, robotics, and transportation.

GPT-3 used few-shot prompts, in which examples of solved tasks are provided as input to the trained model, to fine-tune its performance in a variety of natural language processing tasks. Reynolds and McDonell [37] argue that, with few-shot examples, this system is not learning the task during the run time but rather localize it in its previously learned tasks, and show that zero-shot prompts, that specify the task without examples, can match or even exceed the performance of few-shot prompts. From this conclusion, the authors rethought the role of prompts in LLMs and explored several methods of prompt programming through the lens of natural language. One of the explored methods is direct task specification, which is a zero-shot prompt that tells the model to execute a task that it already knows how to do using a signifier, i.e., a pattern which keys the task intended (like its name or a description), not explaining how to execute the task or expected behaviours. The authors defend that a prompt should constrain the behaviour of the language model to the intended one, not giving it the chance to interpret the prompt the way it is meant to, and also explore the idea of having a prompt that encourages the model to deconstruct the problem for a step-by-step procedure before giving a response, which improves performance in math problems, analogy questions and reasoning over paragraphs. It introduced the idea of metaprompt programming, which uses the language model itself to generate task-specific prompts, saving human investment time.

2.6.1.1 LLMs and Sign language

Lee et al. [38] explore how to use LLMs for Sign Language Translation (text-to-gloss and gloss-to-text) using Vocabulary Sharing by leveraging the high lexical similarity. For this, transformer-based language models are used to model the text-to-gloss and gloss-to-text tasks, using self-attention mechanisms to generate the next token based on the previously generated tokens and the input sequence. Vocabulary Sharing is used to improve the lexical similarity between spoken and sign languages, focusing on glossing and assuming that mapping the representations of the token sequences from glossing to a common latent space with spoken language helps to achieve this improvement. The experiments done proved the efficiency of this approach over other Sign Language Translation models.

2.6.2 Cross-Modal Encoders

A cross-modal task is a type of task that involves more than one modality, such as vision, audio and language, requiring the model to understand the relationship between these different modalities [39]. A cross-modal encoder is a key component of a system that is intended to perform these tasks as it enables it to manage and understand different types of data.

Visual Question Answering (VQA) and image captioning are two examples of a cross-modal task. Visual question answering is when an image is given to the model and the user writes natural language questions about that image for the model to answer. In image captioning the dataset is a combination of images with their respective captions and the model is trained to generate captions for unseen images. Both of these tasks have been extended to be executed with a video (video question answering and video captioning) and all of them rely on convolutional and recurrent neural networks for representing image, video and language.

There are also some variations of VQA and image captioning as benchmark tasks used to pre-train cross-modal models. One of the most used ones is Visual Commonsense Reasoning (VCR), presented by Zellers et al. [40], which extends VQA by not only asking a question about an image but also asking for the reasoning behind the answer that the model chooses. The dataset used for VCR includes combinations of questions, answers and rationales and thousands of movie frames, and as the model receives an image, a question, a set of answers and a set of rationales it has to choose the correct answer and the correct rationale for the given question, as shown in figure 2.7.

BERT (bidirectional encoder representations from transformers), proposed by Devlin et al. [41], is a transformer-based model that uses a self-attention mechanism and position encoding to learn bidirectional language representations. It has two pre-training tasks, masked language modelling (the model is trained to predict previously masked words based on the rest of the sentences) and next sentence prediction (two sentences are provided and the model has to classify if the second sentence is succeeding the first one or if it is just a random sentence), which are solely designed for language, although some

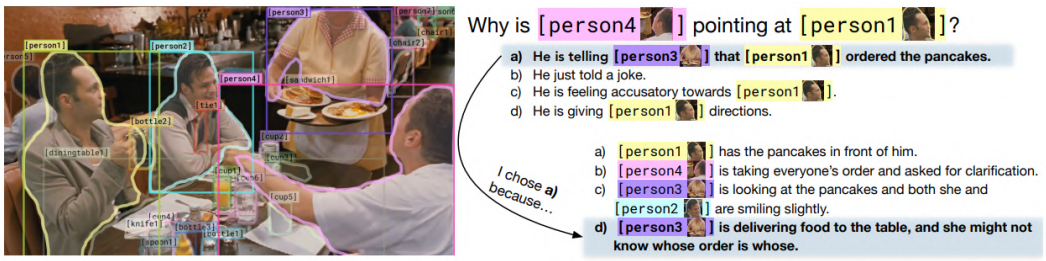


Figure 2.7: In Visual Commonsense Reasoning, an image, a list of regions, a question, a list of answers and a list of rationales are given and the model has to pick the correct answer and the correct rationale [40].

cross-modal models extend these tasks to a cross-modal setting or even create additional ones specifically for this setting.

Masked language modelling is used almost with no exception as it is for linguistic input tokens in cross-modal models, despite the non-sequential nature of vision that difficult the conversion from text to the vision domain. For example, B2T2 (Bounding Boxes in Text Transformer) [42] extends this task by training the model while seeing the image, i.e., it can use the image information to predict the masked words in the text. Next sentence prediction is normally converted to a binary classification task that checks if the image and sentence inputs are semantically matched. Another example is ViLBERT (Lu et al. [43]) which proposes to mask image regions and trains the model to predict the class distribution of these regions.

Some models incorporate feature regression into masked language modelling, and some even propose more pre-training tasks besides the ones performed by BERT. One of these models is LXMERT (Learning Cross-Modality Encoder Representations from Transformers) [44], designed to learn vision-and-language interactions focusing on the representations of an image and its descriptive sentence. This model is composed of three transformer encoders (Figure 2.8): an object relationship encoder that processes the image embeddings, a language encoder for the sentence embeddings, and a cross-modality encoder that aligns and exchanges information between both modalities. It is pre-trained with five different tasks (Figure 2.9): masked cross-modality language modelling, masked object prediction through RoI-feature regression, masked object prediction via detected-label classification, cross-modality matching, and image question answering. These extra pre-training steps allow the model to learn both intra-modality and cross-modality relationships and achieve satisfactory results performing visual question answering.

The network architecture for cross-modal embedding using transformers can be classified into two categories: single-stream models and two-stream models (Figure 2.10). In single-stream models, like VisualBert (Li et al. [45]) and Unicoder-VL (Li et al. [46]), the transformer block is modality-specific and the inputs from both modalities are treated as one, while in two-stream models, like ViLBERT (Lu et al. [43]) and LXMERT (Tan and Bansal [44]), the inputs to each transformer block are inter-modal. Most two-stream

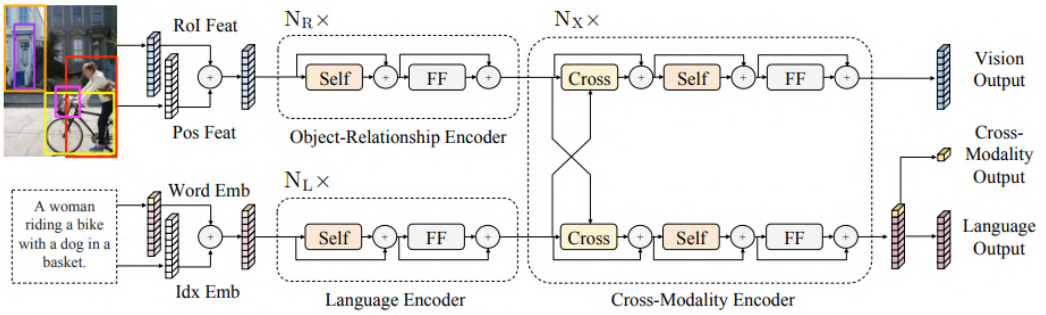


Figure 2.8: The architecture of the LXMERT [44] model. 'Self' and 'Cross' are abbreviations for self-attention sub-layers and cross-attention sub-layers, respectively. 'FF' denotes a feed-forward sub-layer.

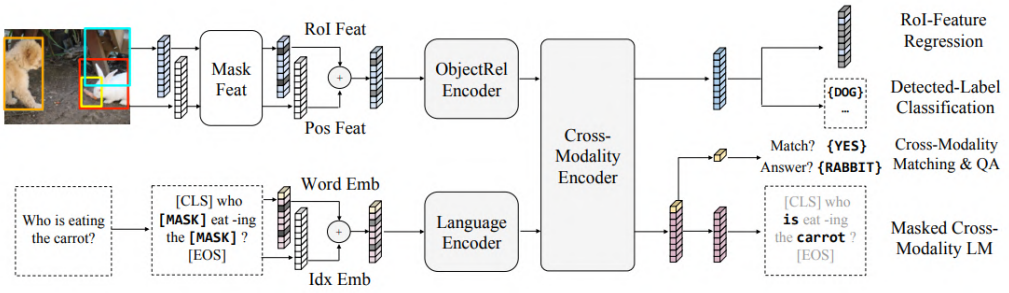


Figure 2.9: The LXMERT [44] pre-training tasks learn the feature representations based on the masked RoI features and word tokens.

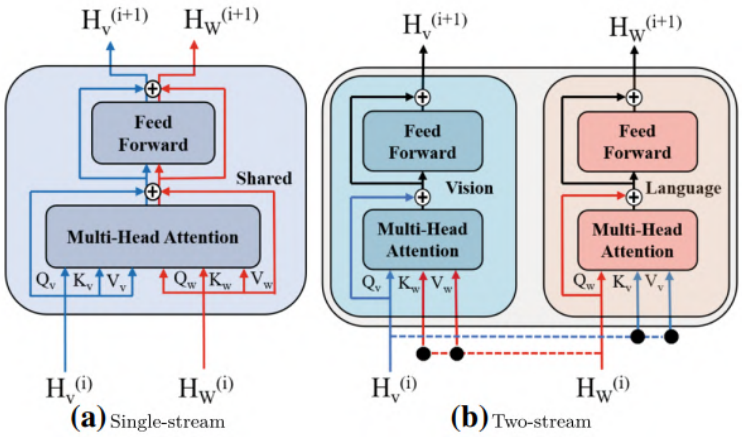


Figure 2.10: Single-stream and two-stream cross-modal transformer blocks. $H_v^{(i)}$ and $H_w^{(i)}$ refer to the embedding of visual and word tokens respectively, output by i-th layer [39].

models incorporate and/or extend an approach similar to the one adopted by ViLBERT, where it uses a co-attention mechanism where inputs from one modality are passed to the transformer block of another modality, allowing for the model to learn features for each modality conditioned on the other.

Nearly all models perform an early fusion of language and video embeddings, where they are concatenated before being fed to cross-modal transformer blocks, which has been

found to outperform late fusion in tasks like VQA and VCR.

In addition to the cross-modal learning mechanism and input format, positional embedding is another factor of variation among the models. Unicoder-VL uses Faster R-CNN to extract image regions along with a 5-D vector, but they use the same position embedding for all image regions, instead of random permutations. In VisualBERT, the position embedding for each visual token is matched to the corresponding input token, whenever the alignments between image regions and the input tokens are available.

2.6.3 Video Captioning

Video Captioning is defined as the generation of a textual description in natural language of the things happening in a video, provided as the input, at a specific point in time, using temporal and space-based methods to produce the result. This task is a highly explored topic and has several applications, from helping visually impaired people to generating a set of instructions from a video.

To execute a video captioning task [47] it is necessary to extract features from the images in the form of words, which are used to form sentences describing the video. Most methods follow a neural net-based approach, which utilizes neural networks for visual and textual data to generate descriptions, combining CNN and Recurrent Neural Networks (RNN), in an encoder-decoder framework, as the video is fed into the CNN to generate the image features which are fed into the RNN in the form of words.

The RNN has two primary methods of feeding data into them. The first method is a bag of words, which is typically used in small models with a limited vocabulary, employing a matrix-driven system. The second method is word embeddings, which represent the higher dimension vocabulary into fewer dimensions using a neural network to simplify computation, generally used when the vocabulary size is large enough. Some approaches use Long Short-Term Memory (LSTM), an updated version of standard RNN, in order to overcome the vanishing gradients problem by allowing the network to learn when to forget previous hidden states and when to update them by integrating memory units, helping the network to retain important information over long periods of time and preventing the loss of information that can occur in traditional RNNs.

This encoder-decoder approach has been widely adopted for video captioning, although it only relies on the forward flow (video to sentence), not considering the backward flow (sentence to video). To make the most of the backward flow, Wang et al. [48] use the concept of dual learning and introduce an encoder-decoder-reconstructor architecture called RecNet (Figure 2.11) to address video captioning. The encoder-decoder generates the semantic representation of each video frame and then produces a sentence description. The reconstructor, implemented using LSTMs, relies on the backward flow and aims to recreate the original video feature sequence based on the hidden state sequence of the decoder. By minimizing the differences between the original and reproduced video features, the reconstructor helps to bridge the semantic gap between the natural language

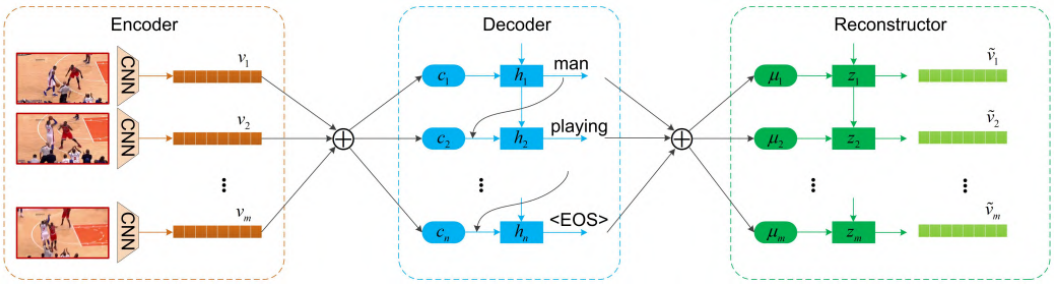


Figure 2.11: RecNet [48] consists of a CNN-based encoder that extracts the semantic representations of the video frames, an LSTM-based decoder that generates natural language for visual content description, and a reconstructor that leverages the backward flow from caption to visual contents to reproduce the frame representations.

captions and video contents.

An attention mechanism can also be included in a video captioning framework to improve its results. Gao et al. [49] propose a video captioning framework named aLSTMs, an attention-based LSTM model with semantic consistency, meaning that the model ensures the generated sentence is semantically consistent with the video content by mapping the generated words and the visual features into a common space. The attention mechanism computes a dynamic weighted sum of local spatial 2D CNN feature vectors at the frame level, representing the most meaningful information at a given time, and allows for the model to focus on the most relevant video features and generate a descriptive sentence. The aLSTMs method has shown competitive results on benchmark datasets for video captioning, outperforming several state-of-the-art methods.

When the language is extracted from the audio of unannotated raw videos, it can be difficult to align language and video as the semantic content of the language and video may not match [39]. For instance, the speaker could be discussing cars, while the video shows the speaker himself. VideoBERT (Sun et al. [50]) explores video captioning and next frame prediction using cooking videos, which have a high probability of visual and linguistic semantics being temporally well-aligned. In this approach, visual tokens are obtained using hierarchical vector quantization applied to video features, and next-sentence classification is extended as alignment classification for visual and linguistic sentences. However, even with cooking videos, the alignment can still be noisy, so the authors address this issue by concatenating neighbouring sentences into a long single sentence and varying the subsampling rate of video tokens, making the model capable of handling different speeds of video playback while ensuring that the alignment between the visual and linguistic elements remains accurate.

2.6.4 Video Compatible Large Language Models

LLMs compatible with inputs of not only text or image but also video add a lot more uses to this type of AI, being able to answer users' requests about the video it has been shown to, such as generating summaries or translations. These LLMs will influence the

VMR technologies, as they can retrieve specific moments that the user asks for. Several developments have already been made in this area.

Video-LLaMA [51] is an in-development model of this type of AI, using two separate branches, one to analyse the video frames and the other for the audio, both being able to translate their input into text queries understandable by LLMs. The one most interesting for this dissertation is the video branch, which the developers call the 'Vision-Language Branch', which starts by extracting features from video frames, then adds temporal information by applying position embeddings to the extracted frames, generates visual query tokens from the frames, and finally converts those tokens to queries compatible with the text inputs of LLMs. To extract the features from the video frames, a pre-trained image encoder is used, which receives a big dataset of short videos with textual descriptions and is trained to generate those text descriptions of the input video. Despite the satisfactory results, Video-LLaMA still has some limitations, like the influence of the video and audio quality on returning accurate responses, the amount of computational resources required to analyse a long video, and sometimes it may generate wrong or irrelevant information.

As mentioned previously, maintaining the dialogue history is essential for the AI agent to have a fluent and realistic conversation with the user. In the cases where there is also a video input, this history must be used in combination with the video information so the agent can give answers more accurately. This is called Video Dialogue and has also been an area thoroughly explored for its importance in the creation of video-compatible LLMs.

For video dialogue, Jin et al. [52] suggests a mechanism of progressive inference which progressively updates query information based on the dialogue history and video content, stopping when the agent thinks that the query information is sufficient and not ambiguous, using a cross-transformer to learn more about the interactions between the video and the text (Figure 2.12). The authors also developed question generation on this system to make it more complete, as this is a good method to train the model and evaluate it, being more difficult than generation answers. After evaluation, the results obtained were satisfactory, outperforming some existing approaches in this area.

2.7 Critical Analysis

It is possible to conclude that there isn't still much work on sign language recognition that includes facial expressions, which is a major flaw in these systems since, as explained, facial expressions are a crucial element in communication through sign language. This dissertation intends to change that.

The explored methods of video processing revealed how important it is to understand how the sign language videos to be processed are structured and how the indexing and the video annotation can be carried out to ease the video moment retrieval task, i.e., the retrieval of the video segments where it is being performed the signs that are searched by the user.

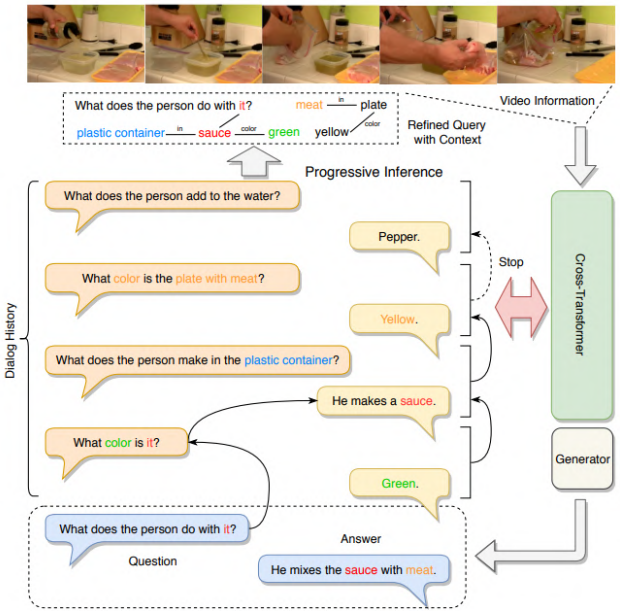


Figure 2.12: Illustration of the Jin et al. [52] video dialog with progressive inference and cross-transformer.

Language models and their capabilities to understand video were also very useful to explore since the usage of both text and video modalities in one system is not easy. After the research, it was possible to gain a better understanding of how these architectures are developed, the possibilities of using these systems and how they can achieve better performance.

After the conclusion of this research, it becomes more feasible to develop this dissertation, having a better knowledge of what is going to be needed and how the best approach and methods can be established to do it.

3

SLVideo: System Architecture and Implementation

In this section, the implementation of SLVideo and its architecture will be presented, discussing all its components, processes, algorithms, the rationale behind the architectural choices, and the dataset of sign language videos used throughout the development. Understanding these elements is crucial to comprehend the development of SLVideo, as it not only provides insight into how this system works but also uncovers the challenges encountered and the solutions that were implemented.

A functioning version of SLVideo is deployed in slvideo.novasearch.org. This system follows the implementation and includes all the functionalities described in this section, such as querying for a facial expression sign, watching the sign language videos from the PSL videos dataset, cycling through the annotated signs and adding new ones.

3.1 Requirements

In order to achieve the objectives established for the development of SLVideo, it was also necessary to define the requirements that this system would have to meet to be considered valuable. These were determined at the beginning of the development process but were continually revised as SLVideo evolved and new objectives were introduced.

Requirements can be either functional or non-functional. The functional requirements define what the system should do, including its features and behaviours, and how the user will interact with the system. In contrast, the non-functional requirements define how the system should perform its functions. The defined requirements are listed in the table 3.1.

3.2 Video and Annotations Formats

To demonstrate and develop SLVideo, we used a dataset with eight hours of annotated video footage featuring people talking in Portuguese Sign Language. The dataset information is shown in table 3.2. A video is a time-ordered sequence of images that depict

Functional Requirements
The system must allow the querying of a facial expression sign or a phrase through a text query;
When using a text query, the system must allow searching using the ground truth;
The system must provide a number of embedding-based search options;
The system must allow the querying of a sign through a video query in the form of a thesaurus;
The system must allow the viewing of the retrieved video segments;
The system must provide information about the retrieved signs;
The system must allow the viewing of the complete sign language videos;
The system must allow the browsing of the annotated facial expression signs of each video;
The system must allow the edition of the annotations regarding the facial expression signs;
The system must allow the deletion of the annotations regarding the facial expression signs;
The system must allow the creation of new annotations regarding the facial expression signs.

Non-Functional Requirements
The system must exhibit modularity, comprising components independent of each other;
The system must exhibit good performance, especially when executing queries;
The system must exhibit extensibility, being easily expanded and updated with new technologies;
The system must exhibit usability, providing a simple and easy user experience;
The system must exhibit flexibility, allowing for easy changes without impacting the entire system.

Table 3.1: Functional and non-functional requirements for SLVideo.

Nº of Videos	**Total Videos Duration**	**Nº of Annotated Videos**	**Total Annotated Videos Duration**	**Total Duration of Facial Expressions Annotations**
504	110:00:40	93	08:08:53	00:54:22

Table 3.2: The dataset properties. The duration information is in the hh:mm:ss format.

the performance of several signs. These signs can be either manual or non-manual and convey a collection of words which compose a phrase.

Each video has annotations associated with it in an XML document, specifically an ELAN Annotation File (EAF) [53]. An EAF file is composed of several annotations regarding the contents of that video. These annotations are divided into tiers (Figure 3.1), each one grouping a different type of annotation regarding the PSL signs and possibly being hierarchically interconnected. The tiers that most interest to SLVideo are the "GLOSA_P1_EXPRESSÃO", which contains all the facial expression signs glosses, its parent tier with the id "GLOSAS_P1", which includes all the signs glosses, and the tier "LP_P1 transcrição livre", which contains all the phrases translated from the signs glosses to the Portuguese spoken language. An example of this three tiers relation is available in figure 3.2.

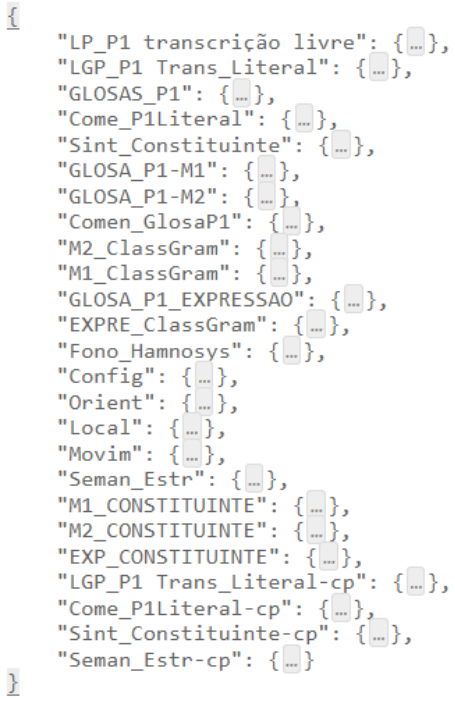


Figure 3.1: The annotations present in an EAF file.

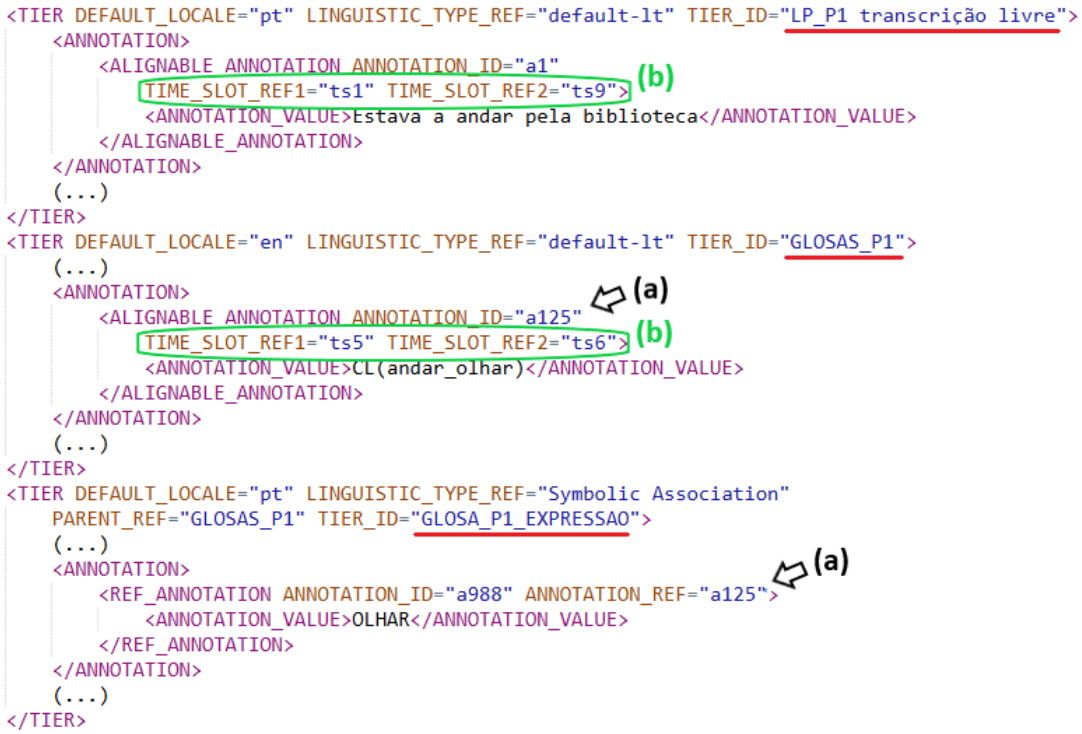


Figure 3.2: An example of the relation of three annotations from three different tiers in an EAF file. (a) We can see that the "GLOSA_P1_EXPRESSAO" annotation references the "GLOSAS_P1" annotation with id a125. (b) As the "time_slot_ref" are incremental and chronologically ordered, we can conclude that the glosses form the "GLOSA_P1_EXPRESSAO" and the "GLOSA_P1" tiers are part of the phrase in the annotation of the "LP_P1 transcrição livre" tier.

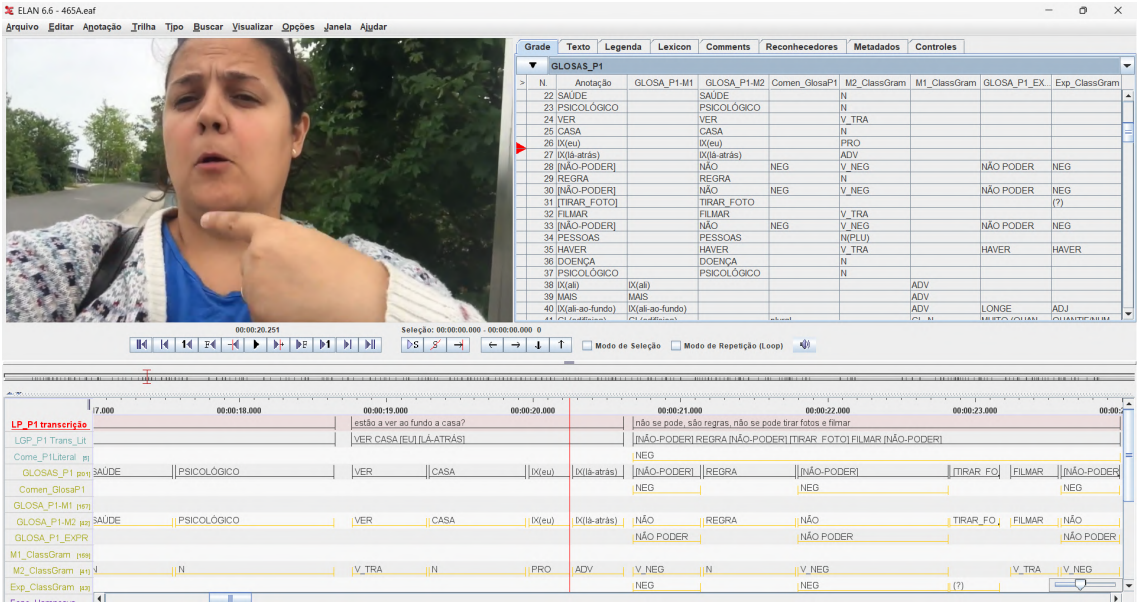


Figure 3.3: ELAN execution screenshot with annotations of a Portuguese Sign Language video.

Alongside the tiers division, these annotations can either be timely aligned to the video, which is the case for the "LP_P1 transcrição livre" and "GLOSAS_P1" tiers, or referring to another annotation, like the "GLOSA_P1_EXPRESSÃO" tier. The timely aligned annotations have a start and end timestamp associated with them and can be a literal translation of the signs executed in the video to a written form in Portuguese or a translation to a gloss annotation in Portuguese. The annotations referring to other ones are defined as symbolic associations, since they give more information about the annotation it is referring to, such as the type of phrase that is being said and the grammatical class of the gloss annotations.

The EAF annotations are designed to be used in ELAN [53], a free software capable of creating complex annotations for video and audio recordings, primarily used in linguistic research, such that *"an annotation can be a sentence, word or gloss, a comment, translation or a description of any feature observed in the media"*[1]. This software was used by Portuguese Sign Language linguistics to create the dataset of annotated PSL videos that are used in the development of SLVideo. A screenshot of ELAN execution with a PSL annotated video can be seen in figure 3.3.

3.3 System Architecture

SLVideo presents a modular architecture, comprising distinct components that complement each other, each assigned a specific task and organized as demonstrated in the file tree of figure 3.4. These tasks include how the video will be indexed and processed, how

[1] https://www.mpi.nl/tools/elan/EAF_Annotation_Format.pdf

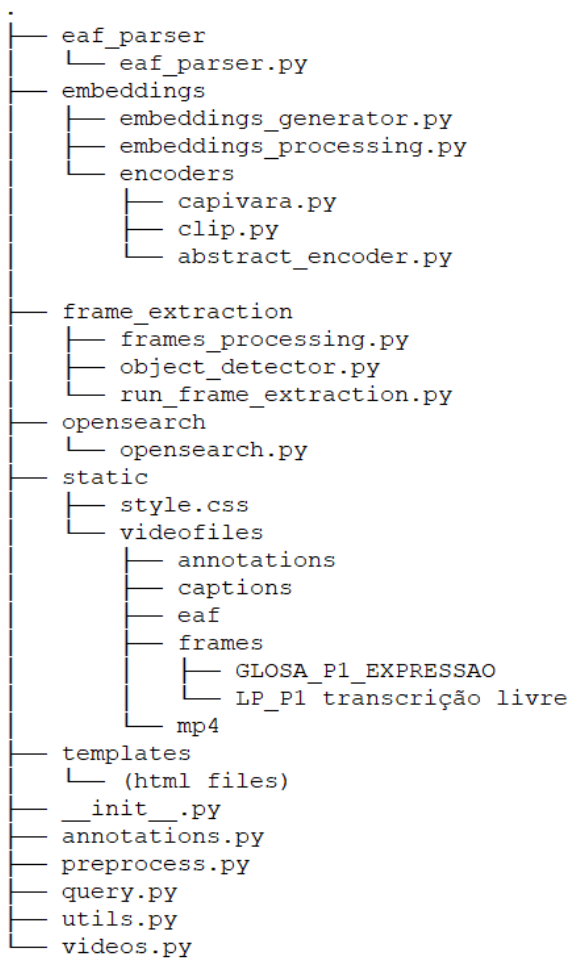


Figure 3.4: SLVideo's file tree.

the annotations will be used, and how the user queries will be interpreted. Its modularity guarantees that this system is efficient, adaptable and flexible for further improvement and expansion.

Figure 3.5 illustrates the relationship between the SLVideo modules involved in the video moment retrieval task, divided into a pre-processing and indexing phase, which is executed when the system starts, and a querying phase, which is executed every time a query is done. The definition of this architecture prior to its implementation is crucial to ensure that all modules work together in an efficient manner to fulfil all the requirements and objectives necessary for the SLVideo system to be effective and valuable.

3.3.1 Annotations Parsing

This is the first step of the pre-processing and indexing phase, where the annotations in the EAF files are the input of a parser that outputs more readable annotation files in the JSON format. This way, the parsed files will only have the relevant information and all the annotations will have a start and end timestamp, contrary to the EAF files where only some annotations had those timestamps.

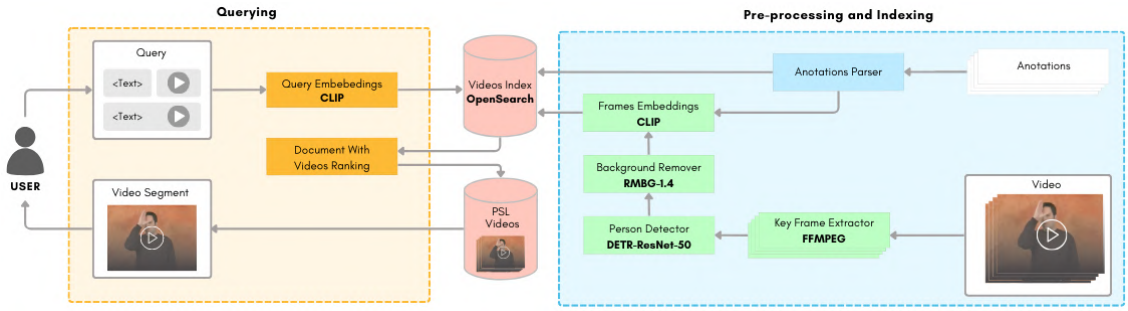


Figure 3.5: SLVideo Moment Retrieval System.

```
"GLOSA_P1_EXPRESSAO": {
  "linguistic_type_ref": "Symbolic Association",
  "tier_id": "GLOSA_P1_EXPRESSAO",
  "annotations": [
    {
      "annotation_id": "a1214",
      "annotation_ref": "a20",
      "value": "ALI",
      "start_time": "31330",
      "end_time": "32668",
      "phrase": "A lebre está ali",
      "user_rating": "4"
    },
    (...)
  ],
  "parent_ref": "GLOSAS_P1"
},
```

Figure 3.6: Example of a parsed "GLOSA_P1_EXPRESSAO" annotation.

This parsing also includes the addition of new information to these annotations, such as by transferring it from another annotation or by creating new fields to be later filled by the user. In the case of the "GLOSA_P1_EXPRESSAO" annotations, the new fields to be added are the start and end timestamps, which corresponds to the ones found in the respective "GLOSA_P1" parent annotation, the phrase from the "LP_P1 transcrição livre" annotation where the "GLOSA_P1_EXPRESSAO" annotation is inserted, and the user rating of the annotation quality, which will be explained later in this chapter. An example of a parsed "GLOSA_P1_EXPRESSAO" annotation is available in figure 3.6.

Additionally, the parsing process incorporates the generation of caption files to be used when watching the sign language videos. These captions will feature complete sentences in the Portuguese spoken language, thus helping non-sign language speakers to understand the PSL videos.

After this parsing, the analysis of the video information is facilitated, not only by simplifying the code but also by being more easily readable directly by the developer.

3.3.2 Video Offline Processing

The second step of the pre-processing and indexing phase is the processing of the Portuguese Sign Language videos. One of the main steps of this process is the extraction of the videos' frames, which is carried out using FFmpeg[2], an open-source software used for video and audio manipulation, receiving the path to the video file and the start and end timestamps as input.

As previously stated, the videos used in the system are a time-ordered sequence of frames. However, not all frames need to be extracted, since that would lead to data redundancy, as frames from the same shot give highly similar information. The only frames that need to be extracted are the keyframes (or I-frames) as a way to only have to analyse relevant frames with different information from each other.

With the parsed annotations file, it is possible to know the start and end timestamps of every gesture with a facial expression, as this information is added to the "GLOSA_P1_EXPRESSAO" tier. The frames within those timestamps are classified as keyframes, so all of them are extracted. Additionally, one frame per "LP_P1 transcrição livre" annotation is also extracted, but, currently, it serves only as an illustration when searching for a phrase using the ground truth, as no further procedure is applied and no embeddings are generated from these frames.

When processing the frames, it's necessary to keep in mind that the main element of the sign language videos is the person performing those signs, with particular emphasis on their hand gestures, body movements and facial expressions. Therefore, every extracted keyframe will be processed using the DETR-ResNet-50 [54] object detection model to detect the person in the frames and return their bounding boxes, cropping it and eliminating unnecessary noise from the frame. Subsequently, the RMBG-1.4[3] image segmentation model is used to remove the background, leaving only the person in the frame.

Through this process of cropping the frames and removing their background, the element taken into account when generating the embeddings is solely the individual performing the PSL signs. By concentrating on hand gestures, body movements and facial expressions, the system achieves better performance.

3.3.2.1 Threading and Batch Processing

The frame extraction, cropping and background removal processes are computationally expensive and can become excessively slow. To address this, threading and batch processing were implemented to accelerate these operations, such that, after testing, we found that the ideal is to have four threads to carry out these three operations concurrently, and batches of 16 frames when executing the cropping and background removal.

These two techniques make the pre-processing stage of SLVideo more efficient. Threading reduces the time required for these tasks by allowing for multiple video segments to

[2] FFmpeg website
[3] RMBG-1.4 documentation

be processed simultaneously, while batch processing avoids processing each frame individually and optimizes resource utilization, ensuring that the hardware is used effectively.

The implementation of threading and batch processing also enhances the system's scalability, such that, as the volume of data grows, the number of threads and the batch size can be updated to maintain optimal performance. With the current settings, the pre-processing speed is increased by 22%.

3.3.3 Embedding Generation

After processing the frames, all of the extracted features of the annotated facial expressions will be later transformed into embedding vectors. These vectors are high-dimensional representations of the data, capturing the semantic meaning and context of the video segments and annotations.

An image and text model receives a frame or annotation as input and generates its embedding vector. Currently, this can be done using one of two CLIP models, the clip-ViT-B-32, which is the image and text model CLIP, and the CAPIVARA[55], which is optimized for texts written in Portuguese. These models were chosen because there are no off-the-shelf models fine-tuned for PSL detection, which makes generating embeddings for these videos a zero-shot task, so clip-ViT-B-32, being a recognized model for its good results, is a suitable option. Because the queries are written in Portuguese, as well as the annotations of the videos, CAPIVARA is also a good model for testing this task.

How the frames are extracted and then processed will significantly influence the quality of the embeddings, as it's important to reduce the noise to a minimum and to focus on the person and the signs they are executing. For this, as mentioned in the previous section 3.3.2, the frames are cropped and their background is removed after being extracted.

There are six different embedding generation methods implemented, each one offering unique advantages:

- **Base Frames Embeddings**: only embeddings from selected frames are generated and then summed together, with the selection being made by calculating a step size and selecting frames at intervals of this step size, ensuring efficient processing while capturing essential features;

- **Average Frames Embeddings**: all frames embeddings are generated and its average is saved as that facial expression embedding, providing a comprehensive representation of the sign executed;

- **Best Frame Embeddings**: only the best frame embedding is saved for each facial expression, comparing the generated embeddings by the norm of the embedding vectors, ensuring high-quality representation by focusing on the most informative frame;

- **Summed Frames Embeddings**: the three previous embeddings are summed together for each annotation to create a single embedding vector, capturing the combined information from these three types of embeddings and enhancing the richness and accuracy of the representation;
- **All Frames Embeddings**: embeddings are generated for all frames of an annotation and then summed together, capturing detailed temporal information;
- **Annotations Embeddings**: embeddings generated from each of the annotations values, i.e., from the facial expression sign glosses.

Each of these options contributes to the system's overall robustness and adaptability, ensuring that it can handle a wide range of queries and video segments effectively, as some queries might deliver better results when using a specific method. This is also useful to us, the developers, as we can test all of these methods and understand which are the most appropriate ones for our objectives.

Generating embeddings from the extracted frames enables the system to support an embedding-based search. This capability allows the users to perform a search using either a text query or a video query, therefore the system has to support these two types of inputs effectively. By accommodating these two distinct input methods, the system enhances its versatility and usability, giving more search options to the user. The same model has to be used when generating the frames/annotations embeddings and the query embedding so that both embeddings are comparable and the search is executed correctly.

3.3.3.1 Encoders Support

Given the rapid advancements in image and text models, SLVideo needs to be designed with extensibility in mind in order to be easily updated by changing its current embedding generator model to a more efficient one, ideally one fine-tuned to sign language recognition. This would improve the accuracy of the querying results and would make this system more valuable and helpful to the users.

Considering this, the architecture of this module follows the strategy pattern [56], which allows for the flexible integration of different encoding models. This design pattern helps the developers to switch between various encoding models without altering the core functionality of the system.

In the UML diagram of the figure 3.7 it is demonstrated that all encoder models implement a common abstract class with standardized methods for generating embeddings, providing a simpler inclusion of a newer model. After defining the new model class, the developer just needs to go to the embedding generator script (Figure 3.4) and change the model that they would like to use.

Additionally, in order to avoid potential incompatibilities in the dependencies of the employed encoder and the cropping and background removal models, the latter are executed in a virtual environment distinct from the remainder of the system. This

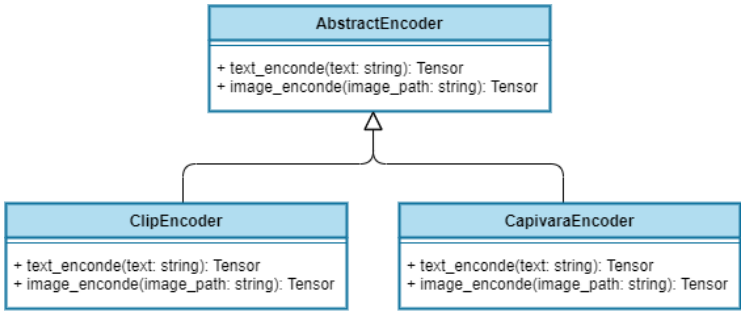


Figure 3.7: Encoders UML diagram.

enhances the extensibility of SLVideo, ensuring that new encoders will not interfere with the frame handling models, simplifying maintenance and facilitating the integration of future updates and improvements.

3.3.4 Indexing Embedding Vectors and Annotations

Some information retrieved from the Portuguese Sign Language videos and annotations needs to be searchable through a query given by the user. This can be achieved by storing the data in documents that can be indexed using OpenSearch[4], an open-source search and analytics suite that allows the user to search, insert, analyse and visualize data. To ensure efficiency, we need to know which of the extracted data is worth indexed, as some are not.

A document is created for each of the annotated facial expression signs and must contain some essential information, including an ID, composed of the video and annotation IDs, as well as the data used to query for a document, which, in this case, are the six types of embeddings generated for that sign. The indexing process involves organizing the documents in a way that allows for fast and accurate retrieval based on the embedding vectors.

It is not necessary to index additional information about the video segments and the annotations, such as the facial expression gloss or the start and end timestamps. This is because such information is not searchable in this context and would only make the searching and indexing processes less efficient. So, when that information is required it can be fetched directly from the parsed annotations files, which is more adequate.

OpenSearch supports several kinds of queries, depending on the embeddings generated and the task at hand, giving different options to the developers to improve their system. Some of these options include filtering information when querying by determining a value for a specific field and filtering the query results later, executing an approximate search, matching specific phrases or words, searching across multiple fields simultaneously, aggregating documents, etc. All of these options were taken into account when developing SLVideo.

[4] OpenSearch website

3.3.5 Query Processing

When searching for a video segment, the queries the user provides can be either text or video, so SLVideo must support both of these inputs. The text queries describe a sign that the user would like to watch in a video segment, and the querying process can be done using the query's plain text or its generated embeddings for an embedding-based search. The video queries are video segments that depict a sign execution, and the querying process includes generating the embeddings from the video frames.

When a user submits a query, the system converts it into an embedding vector using the same model as for the video segments, in order to make them comparable. The search engine then compares this query vector with the indexed vectors, using cosine similarity, to find the most relevant matches.

3.4 User's Functionalities

The defined architecture allows for the implementation of multiple functionalities that are available to the users. The primary functionality is that of video moment retrieval, where the user can search for a particular sign. However, this system also includes a thesaurus search, where signs similar to the one used as a query are retrieved, the visualisation of full sign language videos, and the possibility for users to manipulate the existing annotations and create new ones.

Collectively, these functionalities enhance the user experience, making the system a powerful tool for both learning and teaching sign language.

3.4.1 Search Processes

SLVideo employs two different search approaches, using a text query or a video query, each with its own finality. Currently, text queries are utilised when the user wants to search for particular facial expressions or phrases within the Portuguese Sign Language videos, where video queries enable users to find similar sign language videos based on a selected video segment.

These search processes provide a comprehensive and user-friendly way to explore and analyse PSL videos, thus making SLVideo a valuable asset for both researchers and learners. These functionalities demonstrate the modularity and flexibility of this system, which is capable of handling both text and video queries.

3.4.1.1 Sign Searching and Text Queries

When executing a text-based query to search for a facial expression in the Portuguese Sign Language videos, the user can choose which type of search will be realized depending on whether it will be an embedding-based search or not. It is also possible to search for a phrase that includes the submitted text query, although only using the ground truth.

When querying using the ground truth, the provided text is used to search for correspondences directly in the parsed annotations files, returning all the matching annotations. When performing an embedding-based search, the query embeddings are generated, using the same CLIP model as in the pre-processing phase. These embeddings are then submitted as a query in OpenSearch to perform a k-nearest neighbors (k-NN) search, which returns the ten documents with the highest similarity score with the provided query and its embeddings, which is calculated using cosine similarity.

There are seven options for an embedding-based search, six of them using the embeddings generated in the previous pre-processing phase. The seventh option is a combined search that leverages base frame embeddings, average frames embeddings and best frame embeddings, considering all three types simultaneously and combining their strengths to return the ten best matches.

3.4.1.2 Thesaurus and Video Queries

After performing a text search and selecting a video segment, the user might be interested in knowing signs that are performed similarly to the one it was retrieved to improve their knowledge about sign language. The thesaurus fulfils that request by retrieving the selected video segment's embeddings previously generated and searching, using OpenSearch, for similar videos using those embeddings.

This search is an example of the usage of a video query, where a video segment is processed and submitted to OpenSearch to find the ten matches with the highest similarity score. In this case, the video segment used as a query was already processed in the pre-processing phase, which also helps the search process to be faster and more efficient.

The video segment's embeddings used to perform this search will follow the same method as the one chosen by the user to perform the text search. So, if the user selects the average frames embeddings method, those embeddings are the ones used to perform the video query, except for the combined frames embeddings, the annotations embeddings and searching using the ground truth, which are methods designed solely for text queries.

3.4.2 Collaborative Users' Annotations

ELAN is a powerful tool that provides a full suite of useful features for annotating videos, particularly sign language videos. However, this software does not have collaborative capabilities, posing a challenge for a team that must share the EAF files amongst themselves in order to work together.

SLVideo, being a web application, enables users to collaboratively create, delete, and edit annotations for facial expression signs, as all the users would be handling the same files. When creating or editing annotations, the user can modify the facial expression gloss of the sign, the start and end timestamps of the sign's performance, and the phrase where the facial expression gloss is executed.

This procedure requires the modification of the parsed JSON files, extraction and cropping of new frames, generation of embeddings for these frames, and indexing of the new or altered annotation in OpenSearch. Additionally, it is also necessary to update the EAF files, to reflect the modifications made in SLVideo when accessed through ELAN.

Compared to ELAN, SLVideo handles this process in a much simpler and more basic approach, not having the same level of features. Therefore, for users seeking a complete annotation process, ELAN remains the superior choice, but for those prioritizing collaborative work and requiring less depth in their annotations, SLVideo is the ideal solution.

In addition, users can evaluate the quality of the retrieved video segments and their annotations, whether the segment is timed correctly and whether the annotation is appropriate. This feature not only enhances user interaction but also provides valuable feedback to the developers. By rating the video segments, users directly contribute to the system's evolution, enabling the developers to fine-tune its performance and make necessary adjustments not only to SLVideo but also to the Portuguese Sign Language videos, the annotation dataset used in its development and the generated embeddings.

3.4.3 Sign Language Videos Library

As described in section 3.2, SLVideo comprises several Portuguese Sign Language videos, including storytelling, presentations, two individuals talking to each other, expository videos, and other forms of spoken dialogue in sign language. These videos are the VMR task material, as the user uses them to search for signs, although, SLVideo also allows the visualisation of the complete videos.

This feature is accessible via a library-like format. The videos are accompanied by captions, which enable non-sign language speakers to watch and understand them, as well as a list of the annotated facial expressions signs, enabling the users to view each sign individually.

We believe this is a highly valuable feature for the education of sign language. For example, educators can integrate these videos into their lesson plans, using them as a practical tool to teach sign language, and researchers can employ the videos to study various aspects of sign language communication.

3.5 UX Design Overview

A user interface was developed for SLVideo, which allows the user to search for a facial expression, using the ground truth or seven types of embedding-based search: searching by base frames embeddings, by the average of the frames embeddings, by the best frame embedding, by summed frames embeddings, by all frames embeddings, by combined frames embeddings, or by the annotations' embeddings. A phrase can also be searched, but, for now, only using the ground truth. This interface also includes the thesaurus

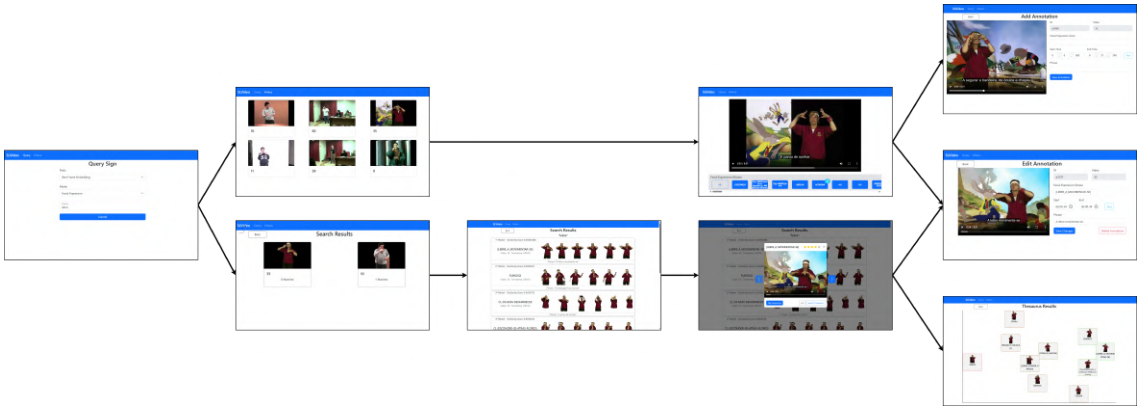


Figure 3.8: UI interactions tree.

search, the navigation on the sign language videos library, watching full videos, and the edition, addition and deletion of annotations. The full workflow can be seen in figure 3.8 and a video demonstration is available in this URL.

SLVideo workflow for the video moment retrieval task follows four steps:

1. The initial page (figure 3.9) allows users to choose whether the search will be embedding-based or through the ground truth, whether it will search for a facial expression or a phrase, and to write and submit the query;

2. Next (figure 3.10), SLVideo presents the videos which have one or more segments corresponding with the query, allowing the user to choose a video;

3. After choosing a video, the users go to the clips page (figure 3.11) where it is shown all the video segments, from the chosen video, ordered by similarity with the query;

4. Once a specific video segment has been selected, a modal (figure 3.12) will open with the video segment. This window allows the user to play the respective sign, navigate to the annotation edition page, do a search in the thesaurus for similar signs, and rate the retrieved video segment using the available rating system;

In the annotation edition page (figure 3.13), the user can change the facial expression gloss of that annotation, the phrase where that gloss is executed, the start and end timestamps, and also delete the annotation.

The thesaurus page (figure 3.14) reflects the similarity between signs by their distance to the queried sign and their border's colour, such as the original sign has a green border, while the least similar sign has a red border. The position of the signs is defined using t-SNE [57], useful for visualizing high-dimensional data by assigning a location to each data point in a two or three-dimensional map.

When clicking on the "videos" link on the navigation bar, it is shown a list of the available videos (figure 3.15), and when clicking on one of those videos, the user is taken

to the video-watching page (figure 3.16) where a list of the annotated facial expression glosses is shown, being possible to click on them and watch the respective segment.

When clicking on a facial expression gloss, a little button will appear that leads to the annotation edition page (figure 3.13), allowing the user to edit that annotation or delete it. At the start of that gloss list, there is a button that allows the user to add a new annotation (figure 3.17), defining its facial expression gloss, the start and end timestamps of the execution of that gloss, and the phrase where that gloss is inserted.

In summary, this interface allows the user to easily query a sign, watch all the available videos and annotated facial expression glosses, edit and rate the existing annotations and add new ones.

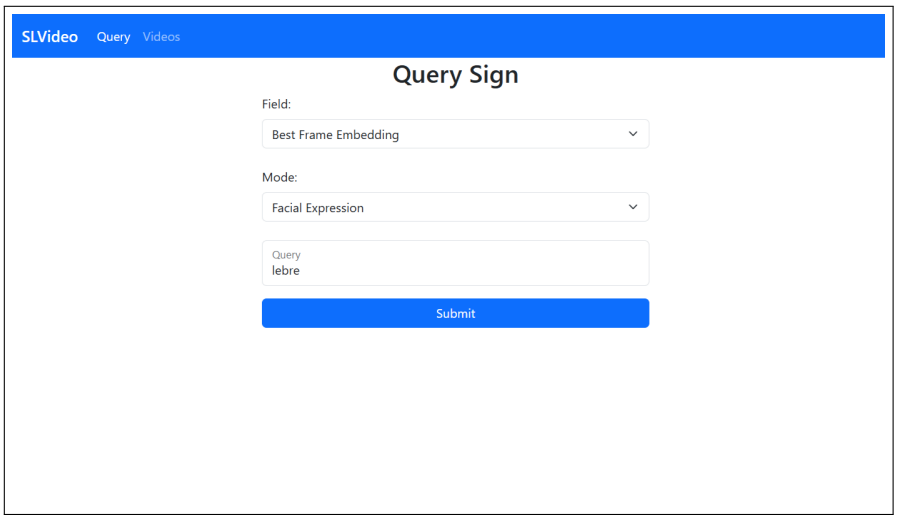


Figure 3.9: Initial page that lets the users search for a sign.

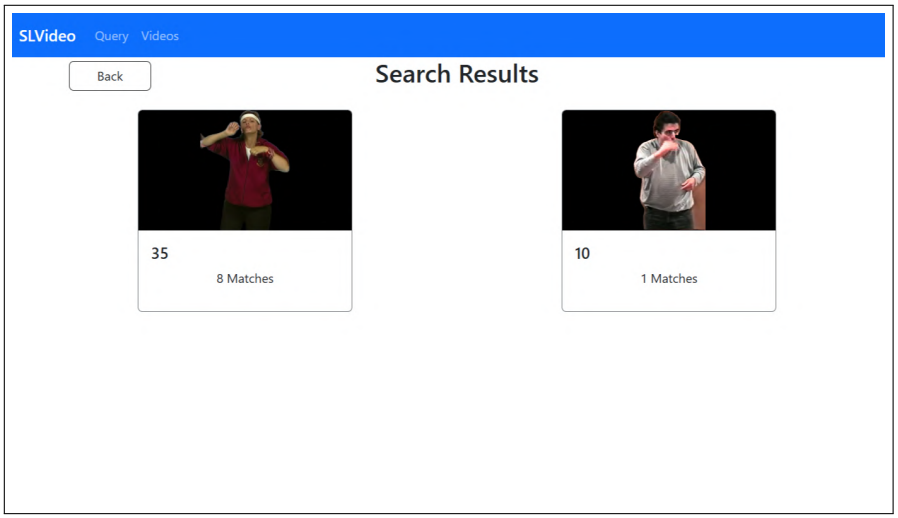


Figure 3.10: The query results displayed by video.

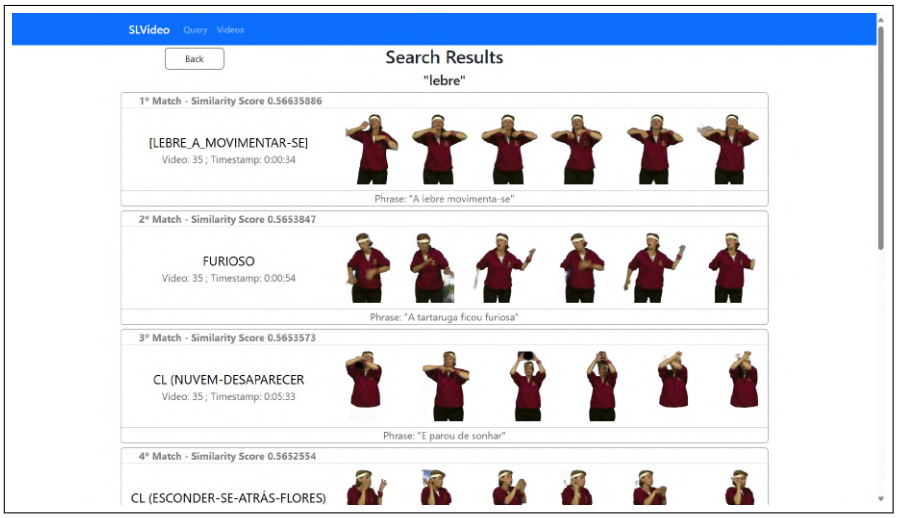


Figure 3.11: The retrieved segments of the chosen video when querying.

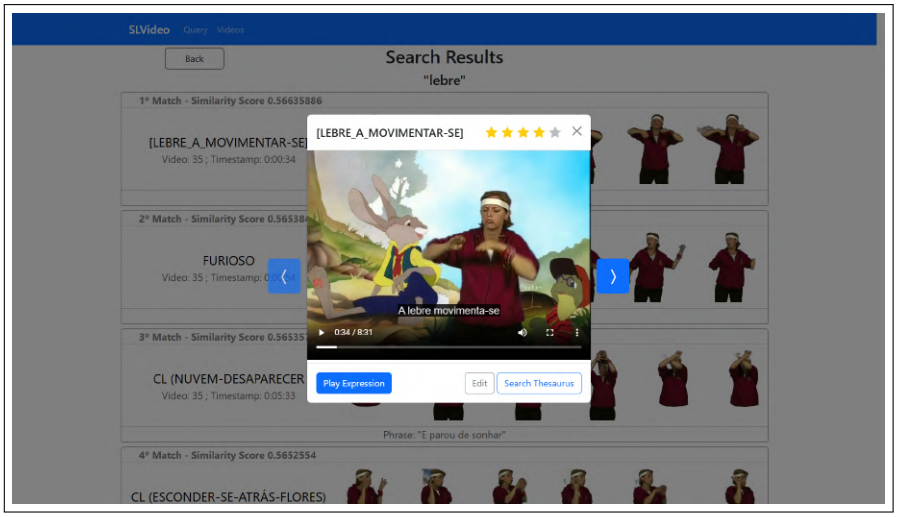


Figure 3.12: Modal with the retrieved video segment.

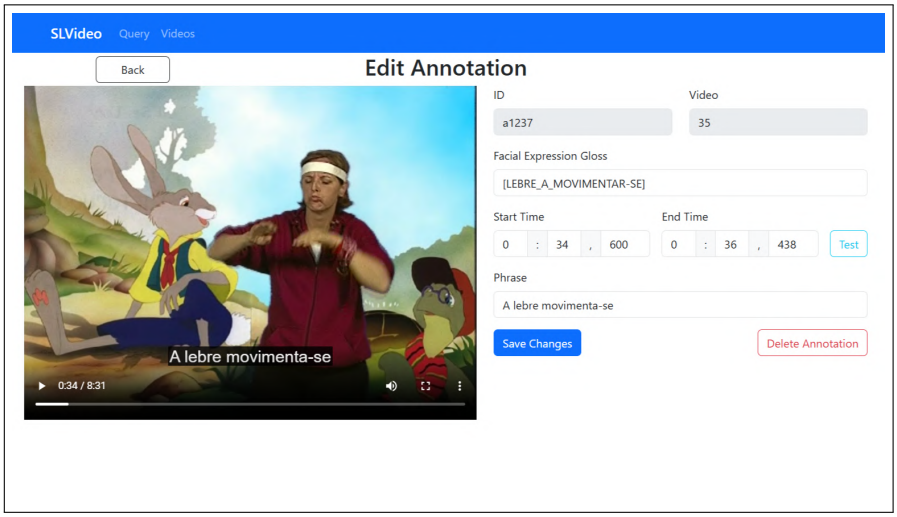


Figure 3.13: The annotation edition page where it's also possible to delete it.

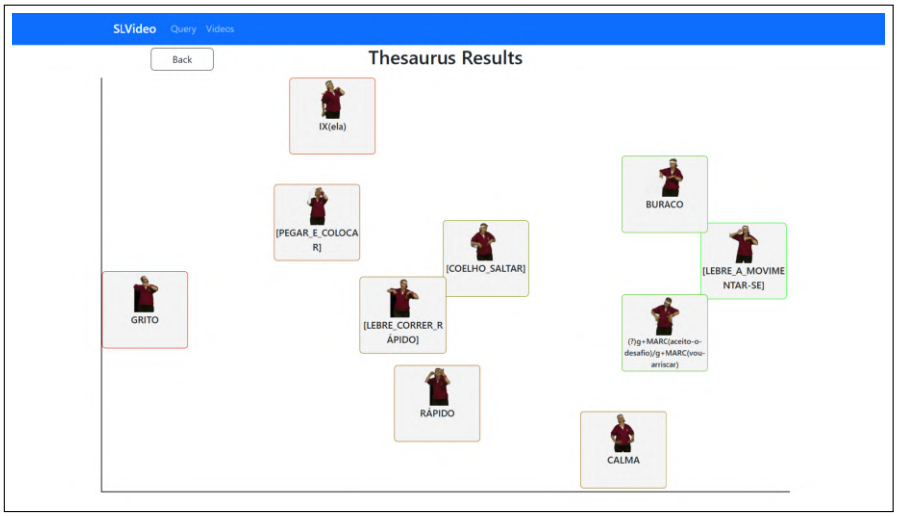


Figure 3.14: Thesaurus page where the user can see similar signs to the one searched.

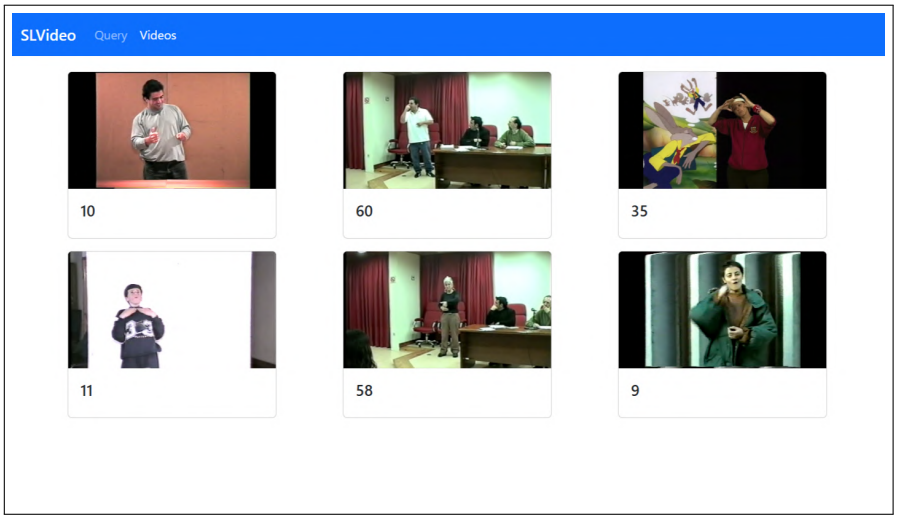


Figure 3.15: The list of available sign language videos in SLVideo.

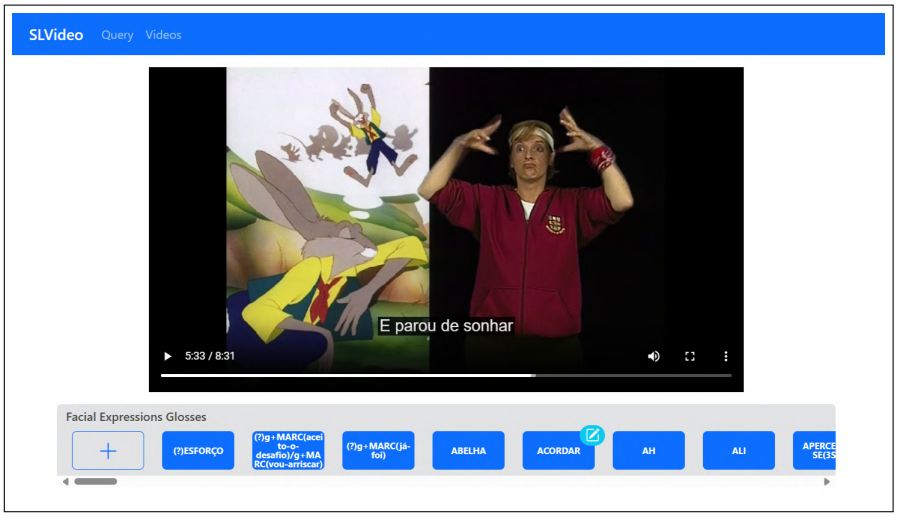


Figure 3.16: The user can watch the video and all the facial expression signs.

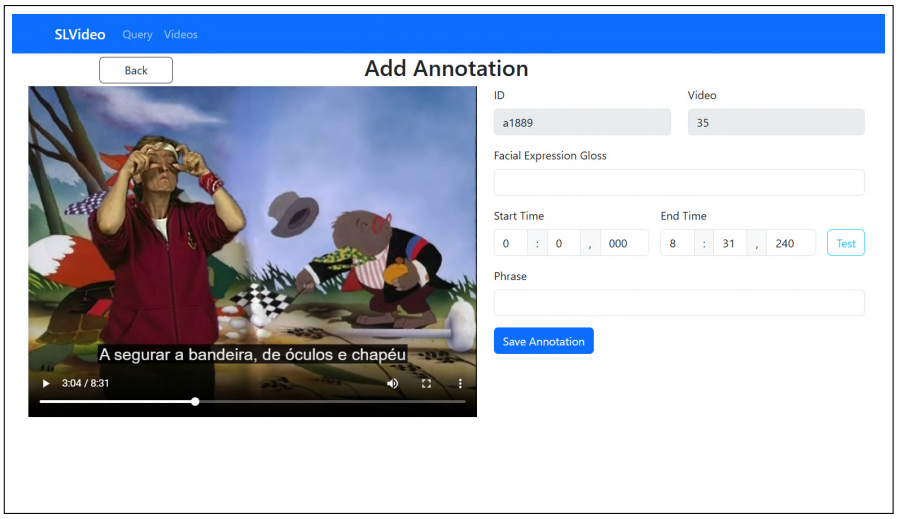


Figure 3.17: The page that allows the user to add a new annotation for a facial expression sign.

4

Evaluation

SLVideo was subjected to constant evaluation throughout its development as several techniques were implemented. This evaluation was crucial to determine the most effective approach, given the necessity of establishing the optimal methodology. This way, it was possible to conclude whether the developed system corresponded to the defined objectives and exhibited satisfactory performance.

In this section, we will present the exhaustive tests conducted to achieve the best version of SLVideo, including how these tests were defined and executed, and our interpretation of the results. These tests were designed not only to demonstrate the efficiency of the current SLVideo architecture but also to illustrate approaches that were not included in the final implementation due to delivering a poorer performance.

4.1 Methodology

As previously stated, the Portuguese Sign Language video dataset is used to test the developed system. SLVideo is evaluated based on the retrieved video segments in response to the user's query, whether it returns the correct ones or not. The relevant video segments are known by retrieving them from the ground truth, i.e. by using the query plain text to search for the desired results directly from the video annotations.

The retrieved video segments will have a similarity score assigned to them by OpenSearch, based on the cosine similarity between the query embeddings and the embeddings of the chosen search method when an embedding-based search is used. When executing a search, the ten video segments with the highest similarity score are returned, although, in the background, the system also retrieves the video segments that match the query directly from the annotation files. A comparison is then made to see how many of the video segments retrieved from the ground truth were also retrieved by OpenSearch, returning values for precision, recall and f1-score, measuring the accuracy of the search.

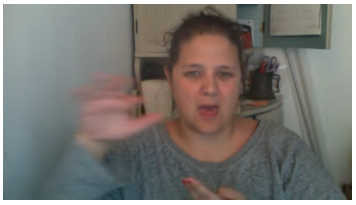
The system's performance is evaluated using precision in most of the tests, as this metric calculates the proportion of relevant retrieved segments. The equation for precision is shown in equation 4.1, where G represents the ground truth results and Q is the query

Video 10

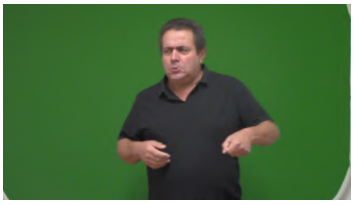
Video 35

Video 408

Video 465A



Video 478



Video 548

Figure 4.1: The six videos used for evaluating SLVideo.

results. We used this metric because it is the most relevant one to our case, as SLVideo retrieves the ten most similar video segments to the query and the test queries have more than ten segments in the dataset. A value of 1.0 indicates that all the retrieved segments correspond to the user's query, whereas a value of 0.1 indicates that one of the ten retrieved segments corresponds to the query.

$$Precision = \frac{|Q \cap G|}{|Q|} \tag{4.1}$$

Using recall, which measures the proportion of relevant video segments retrieved, may lead to misleading results. This is due to the fact that even if all the retrieved segments were relevant, the score would not be perfect if there were more than ten corresponding segments to that query in the dataset. For the same reason, the f1-score is not used in most of this analysis.

In this state, SLVideo has six videos processed and indexed, each with a unique identifier and exhibiting different characteristics, as illustrated in figure 4.1. Two videos have a solid background, with one displaying a lower quality (video 10) than the other (video 548). Similarly, two videos have a complex background, with one displaying a lower quality (video 35) than the other (video 408). One video was recorded by the sign language speaker herself with her phone in "selfie mode" (video 465A), and one was recorded from the chest up (video 478). With these distinct videos, we could draw some conclusions about how the composition of the video affects the final results, not only by performing tests with all videos indexed but also by testing the search execution in each video individually.

The Embedding-based search tests were conducted by querying for the words "muito" (a lot), "não" (no) and "correr" (run) using embedding-based search with Base Frames Embeddings, Average Frames Embeddings, Best Frames Embeddings, Summed Frames

	10	35	408	465A	478	548	Total
Muito	9	4	5	4	14	0	36
Não	0	2	7	4	14	1	28
Correr	0	7	0	0	0	8	15

Table 4.1: Number of times each word appears in each video.

		No Cropping or Background Removal		Cropping and Background Removal	
		clip-ViT-B-32	CAPIVARA	clip-ViT-B-32	CAPIVARA
Muito	Base	0.1	0.3	0.0	0.2
	Average	0.1	0.2	0.0	0.1
	Best	0.2	0.2	0.0	0.1
	Summed	0.1	0.3	0.0	0.2
	All	0.1	0.2	0.0	0.1
	Combined	0.1	0.1	0.0	0.2
Não	Base	0.0	0.0	0.0	0.0
	Average	0.0	0.1	0.0	0.0
	Best	0.0	0.2	0.0	0.2
	Summed	0.0	0.1	0.0	0.0
	All	0.0	0.1	0.0	0.0
	Combined	0.0	0.2	0.0	0.1
Correr	Base	0.0	0.0	0.1	0.3
	Average	0.0	0.0	0.0	0.1
	Best	0.0	0.0	0.2	0.0
	Summed	0.0	0.0	0.0	0.2
	All	0.0	0.0	0.0	0.1
	Combined	0.0	0.3	0.3	0.2

Table 4.2: Results for searching the words "muito" (a lot), "não" (no) and "correr" (run) using the six techniques of frame embedding-based search, two methods of processing the extracted frames and the two models for embedding generation. The metric used in these results is precision, measuring the proportion of relevant retrieved segments.

Embeddings, All Frames Embeddings and Combined Frames Embeddings. These three queries were selected because they are the ones that appear the most in the chosen videos. The table 4.1 illustrates the number of times each word appears in each video.

4.2 Results and Discussion

Following the defined evaluation methodology, we performed different tests to acknowledge the best approach. The two CLIP models used for this evaluation are the same ones used during development, the clip-ViT-B-32 and the CAPIVARA.

As demonstrated in the table 4.2, for both CLIP models, we performed two different tests, one where the extracted frames are given as they are and one where the extracted frames are cropped and the background is removed. An additional test, available in appendix A, where the extracted frames are cropped but the background is not removed, was also performed, but the results are not as relevant as these two tests.

Both CLIP models have similar results, although we can see that CAPIVARA performs better. Additionally, the similarity scores generated by OpenSearch were not satisfying, as for every video segment the score is in an interval of [0.56, 0.59], which is not accurate as the correct video segments should have a higher similarity score. We believe that this is because the models used are not trained specifically for this problem, as they are generic image and text models and are not focused on facial expressions or Portuguese Sign Language interpretation. It's important to note that when using these non-fine-tuned models, this task becomes a zero-shot task, so, despite initial appearances, the results are quite satisfactory for certain cases as it retrieves a good amount of relevant video segments regarding the given query.

We can also see that the results vary a lot depending on how the extracted frames are processed. When no cropping and background removal are done, the frames are given to the model exactly as they were extracted, resulting in embeddings with a lot of noise, particularly when the frames have a complex background with irrelevant information. When cropping and background removal are done there is almost no noise, resulting in cleaner embeddings. These facts may justify the results from both frame extraction techniques, revealing the importance of finding the right balance between noise reduction and information retention to optimize the quality of the extracted frames and the embeddings generated from them.

In theory, the optimal approach would be to crop the frames and remove their background. However, this depends on the quality of the frames and the performance of the models used to perform these tasks, as, in these tests, not executing cropping or background removal seems to lead to the best results. This can be justified by poor cropping and background removal execution, where crucial elements from the frames (like the arms and hands of the individual) were removed, resulting in the loss of very important information and weaker embeddings. Nevertheless, the following tests were executed with the frames cropped and their background removed, as we believe that this approach will be the most efficient one when more advanced models that address this issue are available.

Another tested approach was the utilisation of the YOLOS (tiny-sized) model [58], an object detection model with a design inspired by the DETR model but with some differences. This model was used instead of DETR to crop the extracted frames, followed by the search for the same words as in the initial evaluation. The results of this approach are presented in table 4.3, which did not allow us to conclude which of the two models was better. Consequently, we calculated the medians of the four combinations possible using each of these models with both clip-ViT-B-32 and CAPIVARA, and the results are shown in table 4.4. It was concluded that the optimal approach is to use CAPIVARA with DETR; however, using YOLOS with either CAPIVARA or clip-ViT-B-32 also produces satisfactory results.

After testing the system with all the video annotations indexed, we tested each embedding-based search method with both CLIP models and both DETR and YOLOS

		clip-ViT-B-32	CAPIVARA
Muito	Base	0.0	0.1
	Average	0.0	0.0
	Best	0.0	0.2
	Summed	0.0	0.2
	All	0.0	0.0
	Combined	0.0	0.2
Não	Base	0.0	0.2
	Average	0.0	0.1
	Best	0.0	0.0
	Summed	0.0	0.1
	All	0.0	0.1
	Combined	0.0	0.0
Correr	Base	0.2	0.2
	Average	0.3	0.1
	Best	0.2	0.0
	Summed	0.2	0.2
	All	0.3	0.1
	Combined	0.1	0.0

Table 4.3: Results for searching the words "muito" (a lot), "não" (no) and "correr" (run) using the YOLOS model to crop the frames.

	clip-ViT-B-32	CAPIVARA
DETR	0.03	0.12
YOLOS	0.07	0.10

Table 4.4: Medians of the results of using clip-ViT-B-32 or CAPIVARA with the DETR or YOLOS models.

models, but for each video individually, having only one video indexed at each test. As the number of occurrences of the query words in the videos varies from less than ten to more than ten, we used the f1-score metric for this test, as it balances precision and recall. This helped us verify if the video characteristics would influence the search results, as evidenced by the values displayed in the table 4.5. To simplify the visualization of the results, this table shows a median value for each embedding-search method for each video and embedding model.

The results demonstrate that clip-ViT-B-32 and CAPIVARA give similar outcomes, again, making these tests inconclusive regarding the superiority of either of these two models and emphasising the necessity of the usage of a sign language-specific model yet to be developed. The YOLOS model appears to yield slightly better results; however, DETR also gives satisfactory outcomes. Therefore, in this instance, we gave more weight to the results shown in table 4.4, keeping DETR as the selected model for the final implementation of SLVideo.

Some inferences can be drawn regarding the impact of the video characteristics on the

	clip-ViT-B-32	10	35	408	465A	478	548
DETR	Muito	0.17	0.02	0.16	0.22	0.22	-
	Não	-	0.0	0.32	0.0	0.13	0.0
	Correr	-	0.06	-	-	-	0.4
	CAPIVARA	10	35	408	465A	478	548
	Muito	0.37	0.0	0.11	0.14	0.21	-
	Não	-	0.0	0.1	0.14	0.07	0.0
	Correr	-	0.2	-	-	-	0.67

	clip-ViT-B-32	10	35	408	465A	478	548
YOLOS	Muito	0.17	0.35	0.25	0.17	0.13	-
	Não	-	0.0	0.2	0.0	0.19	0.0
	Correr	-	0.09	-	-	-	0.42
	CAPIVARA	10	35	408	465A	478	548
	Muito	0.35	0.0	0.18	0.34	0.22	-
	Não	-	0.0	0.1	0.36	0.11	0.0
	Correr	-	0.16	-	-	-	0.56

Table 4.5: Results for searching the words "muito" (a lot), "não" (no) and "correr" (run) using the six techniques of frame embedding-based search, the two models for embedding generation and the two models for cropping, but for each video individually. The metric used in these results is a median of the f1-score value for the six embedding-based search methods.

search and the embedding quality. Video 35 exhibits the poorest results, as it is a low-quality video with a complex background, making the cropping and background removal processes less efficient and originating noisy frames with less information. In contrast, videos 10 and 548, which have a solid background, have better results. Based on these observations, we concluded that having better quality videos with solid backgrounds recorded in front of the person leads to better results.

In both previous tests, it can be verified that searching for the word "correr" returns the best results when performing a frame embedding-based search, in comparison to the words "muito" and "não". Given our limited knowledge of PSL, we suspected that this is due to the fact that this sign seems to be performed in a really distinct manner and is clearly separated from the other signs in the sentence, which results in unique embeddings. In contrast, the signs "muito" and "não" are often followed by and associated with other words, such as "muito rápido" (very fast) and "não ser" (not to be), resulting in a rapid succession of signs that are less distinct, producing noisier embeddings.

Another approach that was tested was the usage of an approximate k-NN search[1] method in OpenSearch, which is ideal to improve search speed when using larger datasets, with approximate nearest neighbour (ANN) algorithms from the nmslib [59] and faiss [60] libraries. The outcomes of these tests are shown in table 4.6, which displays only the results for the most effective embedding-based search methods, with the remaining demonstrated

[1] Aproximate k-NN search documentation in OpenSearch

			NMSLIB		FAISS	
			clip-ViT-B-32	CAPIVARA	clip-ViT-B-32	CAPIVARA
DETR	Muito	Best	0.0	0.0	0.0	0.0
		Summed	0.0	0.0	0.0	0.0
		All	0.1	0.1	0.1	0.1
		Combined	0.1	0.0	0.1	0.0
	Não	Best	0.1	0.1	0.1	0.1
		Summed	0.0	0.1	0.0	0.1
		All	0.0	0.0	0.0	0.0
		Combined	0.0	0.2	0.0	0.2
	Correr	Best	0.0	0.0	0.0	0.0
		Summed	0.0	0.0	0.0	0.0
		All	0.0	0.0	0.0	0.0
		Combined	0.2	0.0	0.0	0.0

			NMSLIB		FAISS	
			clip-ViT-B-32	CAPIVARA	clip-ViT-B-32	CAPIVARA
YOLOS	Muito	Best	0.0	0.0	0.0	0.0
		Summed	0.1	0.1	0.1	0.1
		All	0.1	0.1	0.1	0.1
		Combined	0.2	0.0	0.1	0.1
	Não	Best	0.1	0.1	0.1	0.1
		Summed	0.0	0.1	0.0	0.1
		All	0.0	0.0	0.0	0.0
		Combined	0.2	0.3	0.2	0.1
	Correr	Best	0.0	0.0	0.0	0.0
		Summed	0.0	0.0	0.0	0.0
		All	0.0	0.0	0.0	0.0
		Combined	0.0	0.1	0.0	0.0

Table 4.6: Results for searching the words "muito" (a lot), "não" (no) and "correr" (run) using approximate k-NN search with nmslib and faiss engines and space type of L2. All six embedding-based searches were tested, but in this table, we only show the ones that returned the best results.

in the appendix A. These are overall worse than the standard k-NN search methods used in the final implementation. This is because the approximate k-NN search method increases the search processing speed sacrificing accuracy in the search results, which is not desirable as we want the results to be accurate and search speed is not a problem, at least for now.

In addition to the evaluation of the frame embedding-based search approaches, there were also tests conducted for searching using the annotation's embeddings, with the results presented in table 4.7. We can see that clip-ViT-B-32 and CAPIVARA both return excellent results, as all the retrieved video segments correspond to the query. This is because, in this particular test, the embeddings derived from the text query are being compared to those generated from the text of the annotation files, which are more easily matched.

With this exhaustive evaluation, it's possible to confirm that the current SLVideo architecture and implementation is the best approach as it returns the best results with more relevant video segments retrieved.

	clip-ViT-B-32	CAPIVARA
Muito	1.0	1.0
Não	1.0	1.0
Correr	1.0	1.0

Table 4.7: Results for searching the words "muito" (a lot), "não" (no) and "correr" (run) using the annotation's embeddings.

5

Conclusions

This dissertation presented SLVideo, a video moment retrieval framework for sign language videos, focusing on signs where facial expressions have a significant role, recognizing the importance of these expressions in sign language. This chapter will present the main contributions of this system, its limitations and the next steps to take to improve it.

5.1 Contributions

The work done in this dissertation, not just SLVideo but all the research done to develop it, has important contributions not only in the scientific spectrum but also in the social spectrum. Its key contributions are:

- **Scientific research:** To develop SLVideo, extensive research was performed in the areas of sign language linguistics, more specifically on PSL and the roles of facial expressions in sign language, sign language recognition, video processing, video moment retrieval and the application of large language models to videos. This research can serve as a valuable resource for future work in the fields of sign language, artificial intelligence, and video-related studies.

- **Sign Language Video Moment Retrieval:** SLVideo is one of the first video search frameworks to support hand and facial signs with an embedding-based architecture. As such, SLVideo can be seen as a pioneer in this area, showing the importance of the development of this kind of system and the potential it holds for sign language video analysis, setting a precedent for future research and development in this area and providing a robust foundation upon which further advancements can be built.

- **Support for modular sign language encoders:** The two different encoders that were used to extract video embeddings demonstrate the modularity of SLVideo. This design makes the integration of new encoders into the system more easy, like a sign language-specific encoder, providing this system with the flexibility to adapt and evolve with advancements in technology and research.

- **Collaborative tool for annotators of sign language videos:** SLVideo provides its users with the possibility to add new annotations to the available sign language videos, as well as edit or delete the current annotations. It sets itself apart from other video annotation tools by being deployed online, making it a collaborative tool for annotators to work together with the same videos and annotation files.

- **Sign language thesaurus:** SLVideo incorporates a sign language thesaurus to help its users find similar signs to the ones the system retrieves after executing a search. The signs included in the thesaurus are displayed in a very clear manner, facilitating the user's visualisation and comprehension.

- **Access to sign language videos:** Users can watch full sign language videos through SLVideo, as this system provides access to a diverse collection of videos with different characteristics and contextual settings. These videos are accompanied by tools that facilitate sign language learning and research, like the visualisation of singular facial expressions signs' glosses.

- **Deployed application:** Using the Portuguese Sign Language videos dataset, there is already a deployed version of SLVideo ready-to-use, where the user can search for signs through a text query, watch the full videos and facial expressions signs executed in them, and manage the annotations associated to those videos. This deployed version can be accessed in slvideo.novasearch.org.

- **Social contributions:** As exposed in Chapter 1, sign language video moment retrieval software like SLVideo can help deaf or hard-of-hearing people in their lives, easing the communication between them and hearing people, as a hearing person could search for a phrase or word and show the retrieved video to the deaf person. This system can also be used as a learning tool for sign language students by demonstrating the correct execution of signs. Therefore, SLVideo can make an impact on sign language communication and contribute to a more inclusive world for the deaf and hard-of-hearing community.

5.1.1 Paper

This dissertation also contributed to the writing of the paper "SLVideo: A Sign Language Video Moment Retrieval Framework" (see Annex I), demonstrating the architecture of a previous version of SLVideo, its implementation choices, its practical applications, its evaluation and the results derived from it. It also demonstrated a new approach for a model capable of generating more accurate and expressive embeddings for Portuguese Sign Language gestures, considering the entire spectrum of non-manual signals.

This paper will be submitted to an appropriate conference and its webpage is available in novasearch.github.io/SLVideo/.

5.2 Challenges and Limitations

The creation of SLVideo presented numerous obstacles that not only complicated its development but made it more interesting.

A primary challenge was the developers' limited understanding of sign language, particularly PSL. This made system evaluation difficult as it required complete dependence on annotations provided by PSL speakers.

Sign language has a lot of intricate details and aspects that are challenging to capture. The best approach would be to use a sign language-specific transformer to generate the most relevant embeddings possible. Given that the dataset used for the development of this system consists of Portuguese Sign Language videos, and considering the significant differences between each sign language, much like spoken languages, it would be ideal to use a PSL-specific transformer, however, there are no off-the-shelf transformers like that available to use. Because of that, CLIP, a generic transformer for image and text that is not fine-tuned for sign language, was used, leading to a poorer performance and turning this into a zero-shot task.

Another challenge was the insufficient precision of the models used to crop the extracted frames and eliminate their backgrounds. At times, they would produce images with a person with a cropped arm, which is a significant issue when dealing with sign language frames. Additionally, they did not always effectively remove the background, leading to noisy embeddings. In instances where frames contained more than one person sometimes the wrong person would be cropped, regardless of only one of them speaking in sign language, resulting in an entirely irrelevant frame to the task at hand. This imperfection in the models impacted the accuracy of the sign language interpretation.

All of these obstacles highlight the importance of SLVideo and the study developed in this dissertation, its relevance in the area of video processing, artificial intelligence, and systems focused on sign language, and how necessary it is to keep exploring these areas and improving the technologies involved.

5.3 Future Work

SLVideo is already a powerful tool for video moment retrieval with sign language videos, but its limitations, as demonstrated in the previous section, expose that there is still work to do to improve this system.

One of the main focuses is for this system to use a Portuguese Sign Language-specific encoder, in order to better capture the specific nuances and complexities of PSL. This encoder would deliver more accurate results than the generic image and text encoder that is currently being used.

It would be an interesting feature to allow users to submit their own videos as queries to make a more interactive system, although that would require the system to process videos and their frames in run time, which would result in a less efficient search method.

With this in mind, a faster way to extract the frames from the videos, crop them and remove their background would be a major improvement, accelerating also the pre-processing phase of the system.

It's important for SLVideo to follow along the evolution of object detection and image segmentation models, in order to keep updating this system with those new models. When dealing with frames with more than one person, it is crucial to be able to detect which of them is speaking in sign language, so that it would only crop that person. That way, it would have higher quality frames to work with and generate embeddings from, with better cropping and more accurate background removal, with lesser noise and without removing essential parts of the frames.

Including additional categories of signs and glosses would also be a significant improvement to SLVideo, given that this system is primarily focused on signs involving facial expressions. This would allow the management of annotations about those signs and the execution of an embedding-based search for them. After this enhancement, the introduction of more filtering options to the search results and video watching would also be interesting, as it would improve the user experience when searching for a specific sign.

In order to make SLVideo an attractive annotation tool for sign language video annotators, it is crucial to be open to their feedback. Allowing for more detailed annotations and hierarchical relations, as ELAN does, would enhance the value of this system to these annotators, but would also increase its complexity. Therefore, it is important to also maintain a good balance between completeness and usability. One possible approach would be to provide an option for a more simple annotation experience or a more complex one.

By addressing its current limitations and continuously working to enhance its capabilities, SLVideo will remain at the forefront of sign language video moment retrieval technology and an example in this area, as well as in sign language video management and annotation.

Bibliography

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. i).

[2] E. Silva and P. Costa. "Recognition of Non-Manual Expressions in Brazilian Sign Language". In: 2017-06 (cit. on pp. 2, 6, 8, 10).

[3] A. Duarte et al. "Sign Language Video Retrieval with Free-Form Textual Queries". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022) (cit. on pp. 2, 19).

[4] D. F. Moores. "Partners in Progress: The 21st International Congress on Education of the Deaf and the Repudiation of the 1880 Congress of Milan". In: *American Annals of the Deaf* 155.3 (2010), pp. 309–310. DOI: 10.1353/aad.2010.0016 (cit. on p. 2).

[5] H. Carmo and P. Vaz de Carvalho. "PORTUGUESE SIGN LANGUAGE CURRICULUM: past, present and future". In: *Momento - Diálogos em Educação* 31 (2022-07), pp. 327–349. DOI: 10.14295/momento.v31i02.14498 (cit. on pp. 6, 7).

[6] R. I. d. U. C. P. Veritati. "Para além das mãos: elementos para o estudo da expressão facial (EF) em Língua Gestual Portuguesa (LGP)". In: *Veritati - Repositório Institucional da Universidade Católica Portuguesa* (2023). ISSN: 1647-0559. DOI: 10.34632/cadernosdesaude.2011.2812. URL: http://hdl.handle.net/10400.14/12547 (cit. on pp. 6, 8, 9).

[7] M. Amaral, A. Coutinho, and M. Martins. *Para uma gramática da língua gestual portuguesa*. Caminho Linguística. Caminho, 1994. ISBN: 9789722109819. URL: https://books.google.pt/books?id=yZ2PQAAACAAJ (cit. on pp. 6, 7).

[8] C. Viegas et al. *Including Facial Expressions in Contextual Embeddings for Sign Language Generation*. 2022. arXiv: 2202.05383 [cs.CL] (cit. on pp. 6–8, 11).

[9] I. Almeida, L. Coheur, and S. Candeias. "From European Portuguese to Portuguese Sign Language". In: *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language*

Processing for Assistive Technologies. Dresden, Germany: Association for Computational Linguistics, 2015-09, pp. 140–143. DOI: 10.18653/v1/W15-5124. URL: https://aclanthology.org/W15-5124 (cit. on pp. 6, 7).

[10] P. Escudeiro et al. "Virtual Sign – A Real Time Bidirectional Translator of Portuguese Sign Language". In: *Procedia Computer Science* 67 (2015). Proceedings of the 6th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, pp. 252–262. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2015.09.269. URL: https://www.sciencedirect.com/science/article/pii/S1877050915031154 (cit. on pp. 6, 7).

[11] A. B. Baltazar. *Dicionário de Língua Gestual Portuguesa*. Porto Editora, 2010 (cit. on p. 7).

[12] S. Jiang et al. "Skeleton Aware Multi-modal Sign Language Recognition". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2021), pp. 3408–3418. URL: https://api.semanticscholar.org/CorpusID:232341236 (cit. on p. 9).

[13] D. Bragg, K. Rector, and R. E. Ladner. "A User-Powered American Sign Language Dictionary". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '15. Vancouver, BC, Canada: Association for Computing Machinery, 2015, 1837–1848. ISBN: 9781450329224. DOI: 10.1145/2675133.2675226. URL: https://doi.org/10.1145/2675133.2675226 (cit. on p. 10).

[14] A. Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html (cit. on p. 11).

[15] H. Li et al. *CLIPER: A Unified Vision-Language Framework for In-the-Wild Facial Expression Recognition*. 2023. arXiv: 2303.00193 [cs.CV] (cit. on p. 11).

[16] N. Vasconcelos and A. Lippman. "Statistical models of video structure for content analysis and characterization". In: *IEEE Transactions on Image Processing* 9.1 (2000), pp. 3–19. DOI: 10.1109/83.817595 (cit. on p. 12).

[17] X.-S. Hua and M. Wang. "Video Content Structure". In: *Encyclopedia of Database Systems*. Ed. by L. LIU and M. T. ÖZSU. Boston, MA: Springer US, 2009, pp. 3281–3286. ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_1020. URL: https://doi.org/10.1007/978-0-387-39940-9_1020 (cit. on p. 13).

[18] Y. Rui, T. Huang, and S. Mehrotra. "Exploring video structure beyond the shots". In: *Proceedings. IEEE International Conference on Multimedia Computing and Systems (Cat. No.98TB100241)*. 1998, pp. 237–240. DOI: 10.1109/MMCS.1998.693648 (cit. on pp. 13, 14).

[19] C. G. Snoek and M. Worring. "Multimodal Video Indexing: A Review of the State-of-the-art". In: *Multimedia Tools and Applications* 25.1 (2005), pp. 5–35. ISSN: 1573-7721. DOI: 10.1023/B:MTAP.0000046380.27575.a5. URL: https://doi.org/10.1023/B:MTAP.0000046380.27575.a5 (cit. on pp. 13, 16).

[20] A. Podlesnaya and S. Podlesnyy. *Deep Learning Based Semantic Video Indexing and Retrieval*. 2016. arXiv: 1601.07754 [cs.IR] (cit. on p. 14).

[21] M. Y. Kazi Tani et al. "OVIS: ontology video surveillance indexing and retrieval system". In: *International Journal of Multimedia Information Retrieval* 6.4 (2017), pp. 295–316 (cit. on p. 15).

[22] M. Furini. "ViMood: Using social emotions to improve video indexing". In: *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*. 2015, pp. 761–766. DOI: 10.1109/CCNC.2015.7158073 (cit. on p. 15).

[23] J. Huang et al. "Integration of multimodal features for video scene classification based on HMM". In: *1999 IEEE Third Workshop on Multimedia Signal Processing (Cat. No.99TH8451)*. 1999, pp. 53–58. DOI: 10.1109/MMSP.1999.793797 (cit. on p. 16).

[24] A. A. Alatan, A. N. Akansu, and W. Wolf. "Multi-Modal Dialog Scene Detection Using Hidden Markov Models for Content-Based Multimedia Indexing". In: *Multimedia Tools and Applications* 14.2 (2001), pp. 137–151. ISSN: 1573-7721. DOI: 10.1023/A:1011395131992. URL: https://doi.org/10.1023/A:1011395131992 (cit. on p. 16).

[25] K. Q. Lin et al. *UniVTG: Towards Unified Video-Language Temporal Grounding*. 2023. arXiv: 2307.16715 [cs.CV] (cit. on p. 16).

[26] H. Zhang et al. "Video Corpus Moment Retrieval with Contrastive Learning". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '21. Virtual Event, Canada: Association for Computing Machinery, 2021, 685–695. ISBN: 9781450380379. DOI: 10.1145/3404835.3462874. URL: https://doi.org/10.1145/3404835.3462874 (cit. on pp. 16, 17).

[27] J. Dong et al. "Partially Relevant Video Retrieval". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022 (cit. on pp. 17, 18).

[28] A. Diwan, P. Peng, and R. J. Mooney. *Zero-shot Video Moment Retrieval With Off-the-Shelf Models*. 2022. arXiv: 2211.02178 [cs.CV] (cit. on p. 17).

[29] J. B. Roerdink and A. Meijster. "The Watershed Transform: Definitions, Algorithms and Parallelization Strategies". In: (2000), pp. 187–228 (cit. on p. 18).

[30] R. Cui et al. "Video Moment Retrieval from Text Queries via Single Frame Annotation". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2022. DOI: 10.1145/3477495.3532078. URL: https://doi.org/10.1145%2F3477495.3532078 (cit. on p. 18).

[31] G. Wang et al. "Prompt-Based Zero-Shot Video Moment Retrieval". In: *Proceedings of the 30th ACM International Conference on Multimedia*. MM '22. Lisboa, Portugal: Association for Computing Machinery, 2022, 413–421. ISBN: 9781450392037. DOI: 10.1145/3503161.3548004. URL: https://doi.org/10.1145/3503161.3548004 (cit. on pp. 18, 19).

[32] A. Duarte et al. "How2Sign: A Large-Scale Multimodal Dataset for Continuous American Sign Language". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 2735–2744 (cit. on p. 19).

[33] S. Albanie et al. "BSL-1K: Scaling Up Co-articulated Sign Language Recognition Using Mouthing Cues". In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 35–53. ISBN: 978-3-030-58621-8 (cit. on p. 19).

[34] L. Momeni et al. "Watch, read and lookup: learning to spot signs from multiple supervisors". In: *Proceedings of the Asian Conference on Computer Vision (ACCV)*. 2020 (cit. on p. 20).

[35] Y. Cheng et al. "CiCo: Domain-Aware Sign Language Retrieval via Cross-Lingual Contrastive Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 19016–19026 (cit. on p. 20).

[36] G. Varol et al. "Read and Attend: Temporal Localisation in Sign Language Videos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 16857–16866 (cit. on p. 20).

[37] L. Reynolds and K. McDonell. "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI EA '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380959. DOI: 10.1145/3411763.3451760. URL: https://doi.org/10.1145/3411763.3451760 (cit. on p. 21).

[38] H. Lee et al. "Leveraging Large Language Models With Vocabulary Sharing For Sign Language Translation". In: *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. 2023, pp. 1–5. DOI: 10.1109/ICASSPW59220.2023.10193533 (cit. on p. 22).

[39] A. Shin, M. Ishii, and T. Narihira. "Perspectives and Prospects on Transformer Architecture for Cross-Modal Tasks with Language and Vision". In: *International Journal of Computer Vision* 130.2 (2022), pp. 435–454. ISSN: 1573-1405. DOI: 10.1007/s11263-021-01547-8. URL: https://doi.org/10.1007/s11263-021-01547-8 (cit. on pp. 22, 24, 26).

[40] R. Zellers et al. "From Recognition to Cognition: Visual Commonsense Reasoning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 (cit. on pp. 22, 23).

[41] J. Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, 2019-06, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423 (cit. on p. 22).

[42] C. Alberti et al. "Fusion of Detected Objects in Text for Visual Question Answering". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui et al. Hong Kong, China: Association for Computational Linguistics, 2019-11, pp. 2131–2140. DOI: 10.18653/v1/D19-1219. URL: https://aclanthology.org/D19-1219 (cit. on p. 23).

[43] J. Lu et al. "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf (cit. on p. 23).

[44] H. Tan and M. Bansal. "LXMERT: Learning Cross-Modality Encoder Representations from Transformers". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui et al. Hong Kong, China: Association for Computational Linguistics, 2019-11, pp. 5100–5111. DOI: 10.18653/v1/D19-1514. URL: https://aclanthology.org/D19-1514 (cit. on pp. 23, 24).

[45] L. H. Li et al. *VisualBERT: A Simple and Performant Baseline for Vision and Language*. 2019. arXiv: 1908.03557 [cs.CV] (cit. on p. 23).

[46] G. Li et al. "Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.07 (2020), pp. 11336–11344. DOI: 10.1609/aaai.v34i07.6795. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6795 (cit. on p. 23).

[47] V. Jain et al. "Video captioning: a review of theory, techniques and practices". en. In: *Multimedia Tools and Applications* 81.25 (2022), pp. 35619–35653. ISSN: 1573-7721. DOI: 10.1007/s11042-021-11878-w. URL: https://doi.org/10.1007/s11042-021-11878-w (cit. on p. 25).

[48] B. Wang et al. "Reconstruction Network for Video Captioning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018 (cit. on pp. 25, 26).

[49] L. Gao et al. "Video Captioning With Attention-Based LSTM and Semantic Consistency". In: *IEEE Transactions on Multimedia* 19.9 (2017), pp. 2045–2055. DOI: 10.1109/TMM.2017.2729019 (cit. on p. 26).

[50] C. Sun et al. "VideoBERT: A Joint Model for Video and Language Representation Learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019 (cit. on p. 26).

[51] H. Zhang, X. Li, and L. Bing. "Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding". In: *arXiv preprint arXiv:2306.02858* (2023). URL: https://arxiv.org/abs/2306.02858 (cit. on p. 27).

[52] W. Jin et al. "Video Dialog via Progressive Inference and Cross-Transformer". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019-11, pp. 2109–2118. DOI: 10.18653/v1/D19-1217. URL: https://aclanthology.org/D19-1217 (cit. on pp. 27, 28).

[53] Max Planck Institute for Psycholinguistics, The Language Archive. *ELAN (Version 6.7)*. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. 2023. URL: https://archive.mpi.nl/tla/elan (cit. on pp. 30, 32).

[54] N. Carion et al. "End-to-End Object Detection with Transformers". In: *CoRR* abs/2005.12872 (2020). arXiv: 2005.12872. URL: https://arxiv.org/abs/2005.12872 (cit. on p. 35).

[55] G. O. d. Santos et al. "CAPIVARA: Cost-Efficient Approach for Improving Multilingual CLIP Performance on Low-Resource Languages". In: *Workshop on Multi-lingual Representation Learning (MRL), Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2023 (cit. on p. 36).

[56] A. Shvets. "Strategy Design Pattern". In: *Dive Into Design Patterns*. refactoring.guru, 2021. URL: https://refactoring.guru/design-patterns/strategy (cit. on p. 37).

[57] L. van der Maaten and G. Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html (cit. on p. 42).

[58] Y. Fang et al. "You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection". In: *CoRR* abs/2106.00666 (2021). arXiv: 2106.00666. URL: https://arxiv.org/abs/2106.00666 (cit. on p. 50).

[59] L. Boytsov and B. Naidan. "Engineering Efficient and Effective Non-metric Space Library". In: *Similarity Search and Applications - 6th International Conference, SISAP 2013, A Coruña, Spain, October 2-4, 2013, Proceedings*. Ed. by N. R. Brisaboa, O. Pedreira, and P. Zezula. Vol. 8199. Lecture Notes in Computer Science. Springer, 2013, pp. 280–293. DOI: 10.1007/978-3-642-41062-8_28. URL: https://doi.org/10.1007/978-3-642-41062-8_28 (cit. on pp. 52, 66).

[60] M. Douze et al. "The Faiss library". In: (2024). arXiv: 2401.08281 [cs.LG] (cit. on pp. 52, 66).

A

Appendix

A.1 Evaluation Results

We can see that the results vary a lot depending on how the extracted frames are processed, as demonstrated in table A.1. When no cropping and background removal are done, the frames are given to the model exactly as they were extracted, resulting in embeddings with a lot of noise, particularly when the frames have a complex background with irrelevant information. When cropping is done but background removal is not, the noise is decreased, but still present. When cropping and background removal are done there is almost no noise, resulting in cleaner embeddings. These facts may justify the results from each one of the frame extraction techniques, revealing the importance of finding the right balance between noise reduction and information retention to optimize the quality of the extracted frames and the embeddings generated from them.

Another approach that was tested was the usage of an approximate k-NN search method in OpenSearch, which is ideal for larger datasets, with ANN algorithms from the nmslib[59] and faiss[60] libraries. These tests returned the results shown in table 4.6, which are overall worse than the standard k-NN search methods used in the final implementation. The reason behind this is that the approximate k-NN search method increases the search processing speed sacrificing accuracy in the search results, which is not desirable as we want the results to be accurate and search speed is not a problem, at least for now.

		No Cropping or Background Removal		Cropping but no Background Removal		Cropping and Background Removal	
		clip-ViT-B-32	CAPIVARA	clip-ViT-B-32	CAPIVARA	clip-ViT-B-32	CAPIVARA
Muito	Base	0.1	0.3	0.0	0.0	0.0	0.2
	Average	0.1	0.2	0.0	0.0	0.0	0.1
	Best	0.2	0.2	0.0	0.0	0.0	0.1
	Summed	0.1	0.3	0.0	0.0	0.0	0.2
	All	0.1	0.2	0.0	0.0	0.0	0.1
	Combined	0.1	0.1	0.0	0.0	0.0	0.2
Não	Base	0.0	0.0	0.0	0.0	0.0	0.0
	Average	0.0	0.1	0.0	0.0	0.0	0.0
	Best	0.0	0.2	0.0	0.1	0.0	0.2
	Summed	0.0	0.1	0.0	0.1	0.0	0.0
	All	0.0	0.1	0.0	0.0	0.0	0.0
	Combined	0.0	0.2	0.0	0.0	0.0	0.1
Correr	Base	0.0	0.0	0.0	0.0	0.1	0.3
	Average	0.0	0.0	0.0	0.0	0.0	0.1
	Best	0.0	0.0	0.0	0.0	0.2	0.0
	Summed	0.0	0.0	0.0	0.0	0.0	0.2
	All	0.0	0.0	0.0	0.0	0.0	0.1
	Combined	0.0	0.3	0.0	0.0	0.3	0.2

Table A.1: Results for searching the words "muito" (a lot), "não" (no) and "correr" (run) using the seven techniques of frame embedding-based search, the three methods of processing the extracted frames and the two models for embedding generation. The metric used in these results is precision, measuring the proportion of relevant retrieved segments.

			NMSLIB		FAISS	
			clip-ViT-B-32	CAPIVARA	clip-ViT-B-32	CAPIVARA
DETR	Muito	Base	0.1	0.0	0.1	0.0
		Average	0.0	0.0	0.0	0.0
		Best	0.0	0.0	0.0	0.0
		Summed	0.0	0.0	0.0	0.0
		All	0.1	0.1	0.1	0.1
		Combined	0.1	0.0	0.1	0.0
	Não	Base	0.0	0.0	0.0	0.0
		Average	0.0	0.1	0.0	0.1
		Best	0.1	0.1	0.1	0.1
		Summed	0.0	0.1	0.0	0.1
		All	0.0	0.0	0.0	0.0
		Combined	0.0	0.2	0.0	0.2
	Correr	Base	0.0	0.0	0.0	0.0
		Average	0.0	0.0	0.0	0.0
		Best	0.0	0.0	0.0	0.0
		Summed	0.0	0.0	0.0	0.0
		All	0.0	0.0	0.0	0.0
		Combined	0.2	0.0	0.0	0.0

			NMSLIB		FAISS	
			clip-ViT-B-32	CAPIVARA	clip-ViT-B-32	CAPIVARA
YOLOS	Muito	Base	0.0	0.1	0.0	0.1
		Average	0.0	0.0	0.0	0.0
		Best	0.0	0.0	0.0	0.0
		Summed	0.1	0.1	0.1	0.1
		All	0.1	0.1	0.1	0.1
		Combined	0.2	0.0	0.1	0.1
	Não	Base	0.0	0.1	0.0	0.1
		Average	0.0	0.1	0.0	0.1
		Best	0.1	0.1	0.1	0.1
		Summed	0.0	0.1	0.0	0.1
		All	0.0	0.0	0.0	0.0
		Combined	0.2	0.3	0.2	0.1
	Correr	Base	0.0	0.0	0.0	0.0
		Average	0.0	0.0	0.0	0.0
		Best	0.0	0.0	0.0	0.0
		Summed	0.0	0.0	0.0	0.0
		All	0.0	0.0	0.0	0.0
		Combined	0.0	0.1	0.0	0.0

Table A.2: Results for searching the words "Muito" (a lot), "Não" (no) and "Correr" (run) using approximate k-NN search with nmslbi and faiss engines and space type of L2.

I

SLVideo: A Sign Language Video Moment Retrieval Framework

SLVideo: A Sign Language Video Moment Retrieval Framework

Gonçalo Vinagre Martins, Afonso Quinaz, Carla Viegas, Sofia Cavaco, João Magalhães
NOVA School of Science and Technology
Lisboa, Portugal
{gv.martins,a.quinaz}@campus.fct.unl.pt,{s.cavaco,jmag}@fct.unl.pt

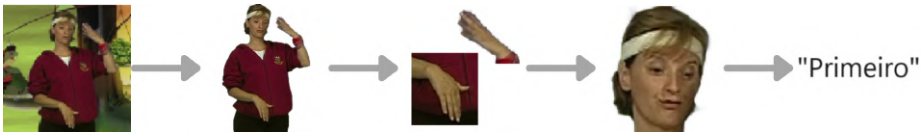


Figure 1: SLVideo is a video moment retrieval framework that is prepared to index and search sign-language content by focusing on the hand, hand-positions and facial signs.

ABSTRACT

Sign Language Recognition has been studied and developed throughout the years to help the deaf and hard-of-hearing people in their day-to-day lives. These technologies leverage manual sign recognition algorithms, however, most of them lack the recognition of facial expressions, which are also an essential part of Sign Language as they allow the speaker to add expressiveness to their dialogue or even change the meaning of certain manual signs. SLVideo is a video moment retrieval software for Sign Language videos with a focus on both hands and facial signs. The system extracts embedding representations for the hand and face signs from video frames to capture the language signs in full. This will then allow the user to search for a specific sign language video segment with text queries, or to search for similar sign language videos. To test this system, a collection of five hours of annotated Sign Language videos is used as the dataset, and the initial results are promising in a zero-shot setting. SLVideo is shown to not only address the problem of searching sign language videos but also supports a Sign Language thesaurus with a search by similarity technique.

Project web page: https://novasearch.github.io/SLVideo/

KEYWORDS

Sign Language Recognition, Video Moment Retrieval.

ACM Reference Format:
Gonçalo Vinagre Martins, Afonso Quinaz, Carla Viegas, Sofia Cavaco, João Magalhães. 2018. SLVideo: A Sign Language Video Moment Retrieval Framework. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In their everyday life, deaf and hard-of-hearing people have to face the challenge of talking to hearing people who don't know sign language. Writing is not always a solution, because the "writing language" is like a whole different language to deaf people, so it works as a second language. The emergence of Sign Language Recognition technologies helps to solve this problem and can support video moment retrieval tasks over sign language videos. Users will then be able to submit a natural language query regarding a video and the system must return the appropriate video segment that corresponds to the query.

A Sign Language Recognition software has to support the recognition of non-manual signs, which include facial expressions and head and shoulder movements, because they are an essential aspect of sign language, due to their grammatical and expressive functions in the dialogue. Facial expressions, specifically, can distinguish what type of phrase is being said, add intensity to what is being said, and even change the meaning of a manual sign [7].

Despite the importance of facial expressions in communication through sign language, most of the approaches for Sign Language Recognition focus on hand gestures and put facial expressions as a low-priority research, as in [2]. This is a concerning fact because the system will lose linguistic information that is present in the facial signs. For a video moment retrieval task with sign language videos, the lack of facial expression recognition could lead to a wrong video segment being returned for a query where the facial expression has an important role.

With SLVideo, we aim to support video moment retrieval systems where the user can search for a sign through a text query and get a set of the relevant video segments corresponding to that query. For this, we are using a five-hour collection of annotated Portuguese Sign Language videos, focusing on the signs that include a facial expression. SLVideo is agnostic of the video encoder and we provide a proof-of-concept with two CLIP [5] encoders for embedding generation and three different techniques for processing the extracted video frames. This system also includes a sign-language thesaurus for the user to search for signs that are similar in terms of gestures and facial signs. Finally, we also support the edition of the video sign-language annotations.

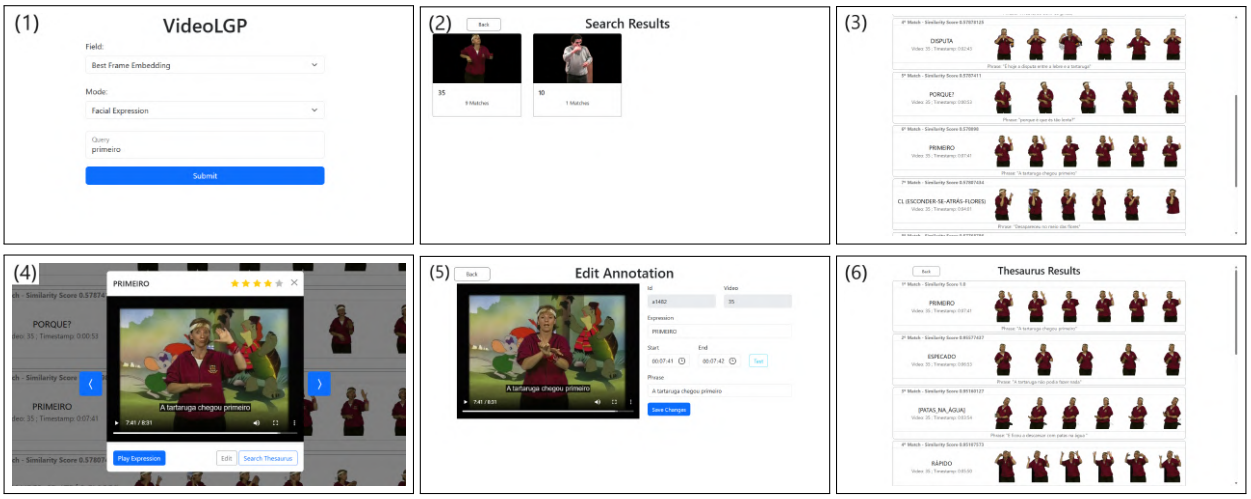


Figure 2: SLVideo workflow follows six steps: (1) the initial page allows users to choose whether the search will be embedding-based or transcripts, whether it will search for a facial expression or a phrase, and to write and submit the query; (2) next, SLVideo presents the videos which have one or more segments corresponding with the query, allowing the user to choose a video and go to the next step (3) where it is shown all the video segments, from the chosen video. After selecting a particular video segment, (4) a modal will open with the video segment. The user can then (5) edit the annotations of that video segment, or (6) go to the sign-language thesaurus.

2 VIDEO AND ANNOTATIONS FORMATS

To demonstrate SLVideo, we used a dataset with more than five hours of video footage featuring people talking in Portuguese Sign Language. Each video is a time-ordered sequence of images that depict the performance of a sign. These signs can be either manual or non-manual and convey a collection of words which compose a phrase.

SLVideo supports videos of any format with or without Sign Language transcriptions/annotations. We use video annotations in the EAF annotation format but other formats are easily supported. The EAF annotation format [4] supports complex annotations for video and audio recordings, primarily used in linguistic research, i.e., *"an annotation can be a sentence, word or gloss, a comment, translation or a description of any feature observed in the media"*. [1]

The EAF file of each video is composed of several annotations, being either timely aligned to the video or it can refer to another annotation. The timely aligned annotations have a start and end timestamp associated with them and can be a literal translation of the signs executed in the video to a written form in Portuguese or a translation to a gloss annotation in Portuguese. The annotations referring to other ones are defined as symbolic associations, since they give more information about the annotation it is referring to, such as the type of phrase that is being said and the grammatical class of the gloss annotations.

3 UX DESIGN OVERVIEW

Currently, there is a ready-to-use user interface as illustrated in figure 2. The system allows the user to search for a facial expression, using the ground truth or four types of embedding-based search: searching by base frames embeddings, by the average of the frames embeddings, by the best frame embedding, or by the annotations' embeddings. A phrase can also be searched, but, for now, only using the ground truth. The search results are divided by videos, with each video showing the number of matches it has with the query, and when selecting a video it will display the video segments, from the best to the worst similarity, so that the user can watch the video segment referring to one of the results by selecting it and can also cycle through the retrieved video segments.

4 SYSTEM ARCHITECTURE

The architecture of SLVideo comprises distinct modules, each assigned with a specific task. These tasks consist of how the video will be indexed and processed, how the annotations will be used, and how the queries will be interpreted.

4.1 Video Processing

The annotations in the EAF file are the input of a parser that outputs a more readable annotations file, in order to have only the relevant information and to have all the annotations have a start and end timestamp, contrary to the EAF file where only some annotations had those timestamps. After this parsing, the video information is analysed more easily.

With the parsed annotations file, it is possible to know the start and end timestamps of every sign with a facial expression. The frames within those timestamps are classified as keyframes, so all of them are extracted. The extraction is done using FFmpeg, an open-source software used for video and audio manipulation. After extracting the frames, a cropping and background removal process will be applied to them to only have the person.

[1] https://www.mpi.nl/tools/elan/EAF_Annotation_Format.pdf

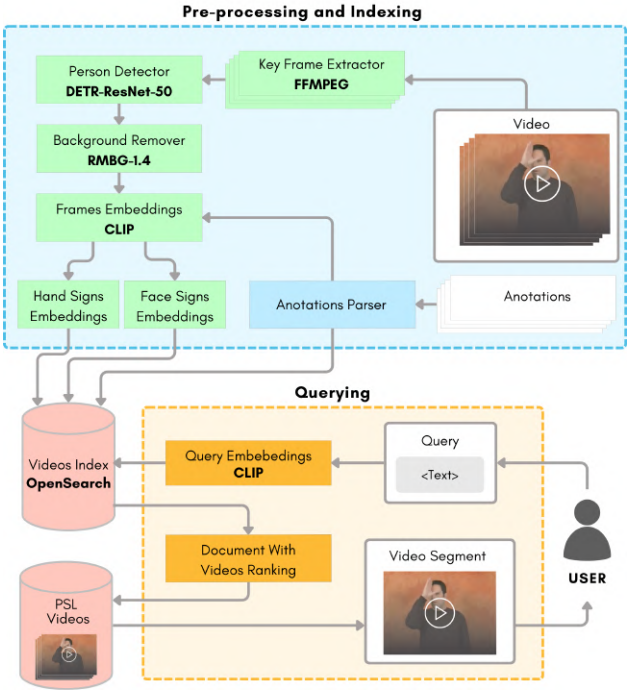


Figure 3: SLVideo Moment Retrieval System.

4.2 Frame Embeddings Generation

The generation of embeddings from the video and from the dataset annotations allows the system to support an embedding-based search. The system was tested using two CLIP models, the clip-ViT-B-32, which is the image and text model CLIP, and the CAPIVARA[6], which is optimized for texts written in Portuguese. Both of these models receive a frame or annotation as input and generate its embeddings. Receiving text as a query, its embeddings can be generated to perform an embedding-based search, using the same CLIP model.

4.3 Facial Signs Embeddings

We extend the capabilities of the existing CLIP model by incorporating a specialized version tailored for Portuguese Sign Language (PSL). This enhanced model leverages the CLIPER[3] framework, which builds on CLIP but is fine-tuned to handle more specific and complex tasks, such as recognizing detailed facial expressions and full-body gestures in PSL.

The standard CLIP model, specifically clip-ViT-B-32, maps text and images to a shared vector space, facilitating the generation of embeddings from both modalities. However, this generic model does not adequately capture the intricacies of PSL, which includes not just manual signs but also essential non-manual signals like facial expressions and body movements.

To address this limitation, we propose the use of a modified CLIPER-like approach. This involves training the model specifically on a dataset comprising annotated PSL videos. This specialized training enables the model to generate more accurate and expressive embeddings for PSL gestures, taking into account the full range of non-manual signals.

In this enhanced setup, the text encoder is also adapted to better align with the linguistic characteristics of PSL. By using datasets tailored for PSL, such as gloss annotations and natural language descriptions of sign language videos, the text encoder can generate more contextually relevant embeddings. This ensures that the text queries match more precisely with the visual content of PSL videos.

4.4 Indexing Embedding Vectors and Annotations

To make the extracted information from the videos searchable through a query given by the user, OpenSearch is a good solution, being an open-source search and analytics suite that allows the user to search, insert, analyse, and visualize data. The information retrieved from the Portuguese Sign Language videos and annotations can be seen as documents that are indexed in OpenSearch, allowing the use of its full-text search capabilities to retrieve these documents.

4.5 Query Processing

The queries the user provides are text-only, and the querying process can be done using the query plain text or its generated embeddings. In the case of text embeddings-based search, the embeddings must be generated using the same model used for generating the frames and annotations embeddings, in order to make them comparable.

4.6 Supported Workflows

SLVideo supports two key workflows, Figure 3: a pre-processing and indexing phase, which is executed when the system starts, and a querying phase, which is executed every time a query is done. It also includes a thesaurus, annotation rating and edition.

4.6.1 Pre-Processing and Indexing. The system starts by parsing the Portuguese Sign Language video annotations in the EAF files to make them more easily analyzed, then uses those annotations' information to extract the keyframes using FFmpeg. Currently, the keyframes are all frames where it is being executed a sign with a facial expression, so all of these are extracted. Every extracted keyframe will be cropped using the DETR-ResNet-50[1] model so that it becomes a smaller image with only the person in the frame, and then the background will be removed using the RMBG-1.4.

With the extracted key frames, embeddings are generated for all of the annotated facial expressions, using the chosen CLIP model, either clip-ViT-B-32 or CAPIVARA. Four types of embeddings are generated: base frames embeddings, where only embeddings from selected frames are generated and then summed together, with the selection being made by calculating a step size and selecting frames at intervals of this step size; average frames embeddings, where all frames embeddings are generated and its average is saved as that facial expression embedding; best frame embeddings, where only the best frame embedding is saved for each facial expression, comparing the generated embeddings by the norm of the embedding vectors; annotations' embeddings, where the embeddings for all the facial expressions annotations' values are generated and saved.

		Primeiro			Rir			Ter		
		Base	Average	Best	Base	Average	Best	Base	Average	Best
No Cropping or Background Removal	clip-ViT-B-32	0.14	0.29	0.14	0.0	0.0	0.0	0.33	0.0	0.33
	CAPIVARA	0.0	0.0	0.0	0.75	1.0	0.57	0.33	0.33	0.33
Cropping but no Background Removal	clip-ViT-B-32	0.14	0.14	0.43	0.25	0.25	0.25	0.33	0.33	0.33
	CAPIVARA	0.0	0.0	0.0	0.5	0.5	0.25	0.0	0.0	0.0
Cropping and Background Removal	clip-ViT-B-32	0.43	0.43	0.29	0.0	0.0	0.0	0.33	0.0	0.33
	CAPIVARA	0.0	0.0	0.0	0.0	0.25	0.0	0.0	0.0	0.0
Annotations	clip-ViT-B-32		1.0			0.5			0.33	
	CAPIVARA		1.0			1.0			1.0	

Table 1: Results for searching for the words "Primeiro" (first), "Rir" (laugh) and "Ter" (have) using the three techniques of frame embedding-based search, the three methods of processing the extracted frames and the annotations embeddings search. The metric used in these results is Recall, measuring the proportion of relevant video segments that were retrieved.

All of this information will be indexed in a document, to be used by OpenSearch, composed of the video and annotation IDs and the four types of embeddings generated for that facial expression.

4.6.2 Querying. The queries created by the user will be text-based, searching for a facial expression or phrase in the Portuguese Sign Language videos. When querying using the ground truth, the provided text is used to search for correspondences in the annotations files, returning all the matching annotations.

When performing an embedding-based search, the query embeddings are generated, also using CLIP, and they are used to search in OpenSearch for the corresponding document, returning the ten matches with the best similarity score with the provided query and its embeddings, which is calculated using cosine similarity.

4.6.3 Thesaurus and User's Annotations. Despite the video moment retrieval task being the most focused task of this system, SLVideo also includes a thesaurus and the ability for the users to edit and evaluate the existing annotations.

When selecting a video segment, the user might be interested to know signs that are performed similarly to the one it was retrieved. The thesaurus fulfills that request, by retrieving the selected video segment's embeddings previously generated and searching, using OpenSearch, for similar videos using those embeddings.

It is also useful to constantly improve the dataset to enhance the CLIP model. This can be done by allowing the users to rate the retrieved video segments and also edit the existing annotations. To guarantee the rating and edition reliability, this would only be available to professional annotators and sign language speakers.

5 EXPERIENCES AND RESULTS

This system is evaluated based on the retrieved video segments regarding the query given by the user, whether it returns the correct video segments or not, being recall the most important metric, as it measures the proportion of relevant video segments that are retrieved. The relevant video segments are known by retrieving them from the ground truth, that is by using the query plain text to search for the desired results directly from the video annotations.

When the search is embedding-based, the retrieved video segments will have associated with them a similarity score based on the cosine similarity between the query embeddings and the chosen search method embeddings. Those similarity scores were not satisfying, as for every video segment the score is in an interval of [0.56, 0.59], which is not accurate as the correct video segments should have a higher similarity score.

As demonstrated in table 1, to test the system we queried for the words "Primeiro" (first), "Rir" (laugh) and "Ter" (have) using embedding-based search with Base Frames Embeddings, Average Frames Embeddings and Best Frames Embeddings. For both CLIP models, we did three different tests, one where the extracted frames are given as it is, one where the extracted frames are cropped to have only the person, and one where the extracted frames are cropped and the background is removed. We also tested the usage of Annotations' Embeddings.

We can see that the results vary a lot depending on how the extracted frames are processed and that both CLIP models have similar results. We believe that this is because the used models are not trained specifically for this problem, as they are a generic image and text model and are not focused on facial expression or Portuguese Sign Language interpretation. It's important to note that when using these non-fine-tuned models, this task becomes a zero-shot task, so, despite initial appearances, the results are quite satisfactory for certain cases as it retrieves a good amount of relevant video segments regarding the given query. As for searching using the Annotation's Embeddings, we can see that CAPIVARA retrieve better results than clip-ViT-B-32, as it is a model focused on the Portuguese language.

6 CONCLUSION

In this paper, we proposed SLVideo, a system to search Sign Language content. The key contributions are:

- **Sign-Language Video Moment Retrieval:** SLVideo is one of the first video search frameworks to support hand and facial signs with an embedding-based architecture.
- **Support for modular Sign-Language Encoders:** The two different encoders that we used to extract video embeddings, demonstrate the modularity of SLVideo.
- **Proof of concept demonstration:** Using a dataset of annotated videos, the system allows users to search for specific signs through text queries.

Future improvements will focus on enhancing model accuracy to better support communication for deaf and hard-of-hearing individuals.

REFERENCES

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. *CoRR* abs/2005.12872 (2020). arXiv:2005.12872 https://arxiv.org/abs/2005.12872

[2] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. 2023. CiCo: Domain-Aware Sign Language Retrieval via Cross-Lingual Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19016–19026.

[3] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. 2023. CLIPER: A Unified Vision-Language Framework for In-the-Wild Facial Expression Recognition. arXiv:2303.00193 [cs.CV]

[4] Max Planck Institute for Psycholinguistics, The Language Archive. 2023. *ELAN (Version 6.7)*. https://archive.mpi.nl/tla/elan Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[6] Gabriel O. dos Santos, Diego A. B. Moreira, Alef I. Ferreira, Jhessica Silva, Luiz Pereira, Pedro Bueno, Thiago Sousa, Helena Maia, Nádia da Silva, Esther Colombini, Helio Pedrini, and Sandra Avila. 2023. CAPIVARA: Cost-Efficient Approach for Improving Multilingual CLIP Performance on Low-Resource Languages. In *Workshop on Multi-lingual Representation Learning (MRL), Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[7] Emely Silva and Paula Costa. 2017. Recognition of Non-Manual Expressions in Brazilian Sign Language. *12th IEEE International Conference on Automatic Face and Gesture Recognition.*