A 2-D Wrist Motion Based Sign Language Video Summarization

Evangelos G. Sartinas[1]
sartinas@ceid.upatras.gr
Emmanouil Z. Psarakis[1]
psarakis@ceid.upatras.gr
Klimis Antzakas[2]
k.antzakas@upatras.gr
Dimitrios I. Kosmopoulos[1]
dkosmo@upatras.gr

[1] Dept. of Computer Engineering & Informatics
University of Patras
Greece
[2] Dept. of Primary Education
University of Patras
Greece

Abstract

In this paper we present a keyframe extraction scheme based on the wrist motion using differential geometry. More specifically, the time (t)-parameterized Frennet-Serret frame for tracking the signer's wrist is used and the curvature of the trajectory, is proposed for the identification of the Sign Language (SL) video keyframes. Specifically, a video frame is characterized as keyframe if on that time instance the t-parameterized curvature function attains a maximum value. Finally, in order to properly define the wrist 2-D motion model, a skeleton tracker is used. The proposed scheme is adaptable, i.e., the number of extracted keyframes varies according to the complexity of the signs, while preserving the semantic content. This in turn makes it attractive for applications like video-calling. Its performance in terms of the achieved compression and intelligibility ratios was evaluated on a ground-truth sequence and outperformed its s-parameterized counterpart (s is the arc length); it also outperformed a moment-based SL summarization technique. Furthermore, the proposed scheme was experimentally evaluated on a dataset containing 5500 signs by SL specialists with very promising results. Finally, the proposed keyframe extraction was evaluated against the aforementioned techniques on the same dataset via the use of a GRU neural network on the gloss classification problem; its superior accuracy in identifying the gloss meaning was confirmed.

1 Introduction

The Sign Languages (SLs) are typically the native languages of the Deaf and of many of the hard-of-hearing (HoH). Due to their poor experiences in spoken or written languages the Deaf typically prefer using SLs to reading or writing text [18]. Nowadays video-capturing devices are ubiquitous and play important role in the communication and education of the Deaf. A method to summarize SL videos, without sacrificing the semantics of the performed signs would offer significant benefits, especially in applications such as communication over low-bandwidth networks, or content browsing.

In the past, several general-purpose video summarization methods were presented (e.g., [22], [13], [24]); however those are not applicable in SL videos, since they treat the video

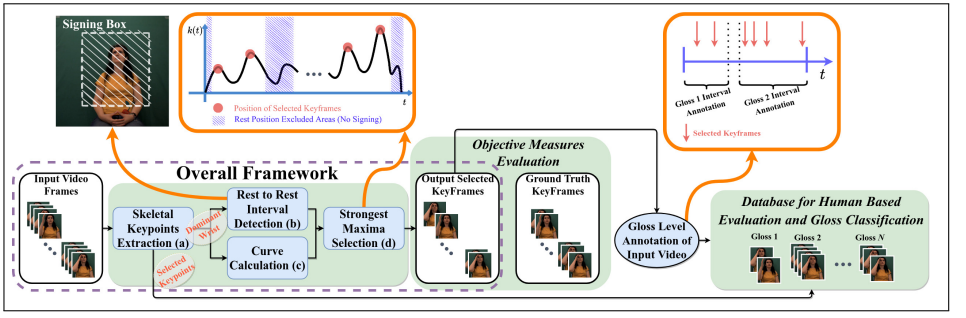


Figure 1: The proposed overall summarization framework (dotted box). Its output is used for the objective measures evaluation (Section 5.2), by comparing it with the ground truth keyframes, and the creation of the database for human based evaluation (Section 5.3) and gloss classification (Section 5.4)

frames holistically, while in the SL videos only some very specific regions are important for interpretation, while the rest of the frame is actually irrelevant. Such regions are associated to specific parts of the human body (mainly hands and face); this fact is in stark contrast to the holistic summarization methods.

In some cases, the use of summarization schemes is necessary for solving more complicated problems, such as SL translation [14]. The key assumption behind this is that if the summary can capture the semantic essence of a video then it can be used for solving the translation problem. The proposed method is expected to impact indirectly the analysis of SL, by focusing on the frames where the most important information exists. This may introduce huge savings in the recognition processes, e.g., by keeping only a small fraction of the data as input to a classifier. It also may impact the research dealing with the efficient transmission of SL videos for low-bandwidth networks.

Signers use multiple channels to convey information that can be grouped under two main categories, namely manual (that are related to the hands and their motions) and non-manual (mainly facial expressions and body pose) features. Sign phonology apart from non-manual features, includes the description of the handshape (hand formation), the movement, the location and the orientation (palm and fingers). In relation to the movement, [7] notes that the wrist moves through space in order to achieve a change of location. Furthermore, [19] refer to the sonority of syllables as the ability of a sign to be perceived at greater distance. Therefore, joints closer to the body are considered to be higher in the rank of sonority. [9] proposes a sonority hierarchy as follows: Shoulder - elbow - wrist - base joints - non-base joints. This hierarchy indicates the importance of the wrist for the perception of the movement of a sign if we take into account that it integrates the motion of shoulder and elbow.

In this work we contribute by (a) introducing a method that does not require any form of training for efficient summarization of SL videos, based on wrist motion, that preserves their lexical meaning and outperforms all known methods and (b) by developing a dataset of continuous SL videos for the evaluation of summarization methods along with its ground-truth keyframes and gloss annotation; to our best knowledge there is no such dataset publicly available.

In Fig. 1 the overall framework of summarizing SL video, which works for every video frame as follows: the skeleton tracker extracts among other keypoints the Signer's dominant

wrist position (a) which is used in turn for the detection of the signing intervals (b) and the calculation of the curve describing the importance of every frame (c). Finally, the important frames are selected using the positions where the selection curve attains its strongest maxima (d).

The rest of the paper is structured as follows: section 2 presents the prior work. In section 3 we formulate the problem, while in section 4 we present the proposed methodology, which is followed by the experimental results in section 5. Finally, section 6 concludes this paper.

2 Related Work

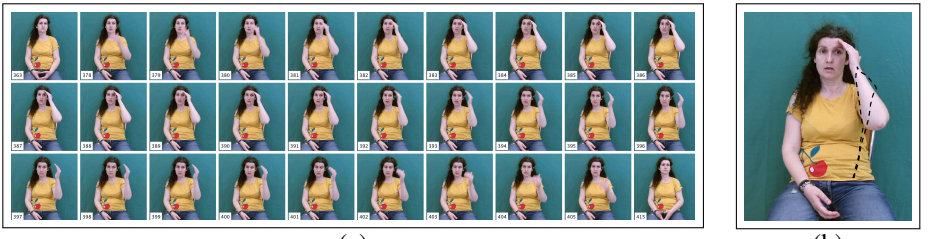
A first line of research deals with the extraction of keyframes from video content, also called "static summary" (in contrast to "dynamic summary" that extracts short videos). Many initial approaches used low-level features such as the color or motion histograms (e.g., [24]), SIFT/SURF (e.g., [13]), or more recently features from pretrained CNNs [22]. Then the keyframes are typically extracted using entropy (e.g., [3]) or clustering methods (e.g., [5]). Such methods mainly use the structural and not the semantic information in the video; however, they count on the fact that the changes in the structural frame data (objects) may be associated to semantic changes, which quite often is true. Some later approaches try to identify the semantic events, which are of importance, like in sports, e.g., [2] or video surveillance, e.g., [20]. To this end, objects may be identified and tracked.

There have been reported supervised methods, which assume human annotations of keyframes in training videos, and seek to optimize the frame selection by minimizing loss with respect to this ground truth. In [28], two LSTMs are used to select keyframes, by minimizing the cross-entropy loss on annotated ground-truth keyframes with an additional objective based on determinantal point process (DPP) to ensure diversity of the selected frames.

Some of the most recent works in unsupervised summarization exploit the auto-encoder architecture combined with recurrent networks such as the LSTM. In [26] the auto-encoder is trained using a proposed shrinking exponential loss function that makes it robust to noise in the web-crawled training data, and is configured with bidirectional LSTM cells to better model the temporal structure of the highlight segments. In [17] the summarizer is an auto-encoder LSTM network aimed at, first, selecting video frames, and then decoding the obtained summarization for reconstructing the input video.

Although, the above mentioned techniques are widely used for the video summarization task, their applicability in SL videos is not straightforward. Most of these techniques relied on the segmentation of the video into shots [17, 21, 28], which is usually done using dynamic or static temporal difference [27] or color histogram similarity among the keyframes [10]. However SL summarization scenario, due to the static scene and the occurrence of abrupt motion, is different which means that the temporal, or color histogram difference, doesn't reflect the similarity between frames while the meaning is drastically changed between consecutive frames.

Another line of research aims to find those tradeoffs for transmission of SL videos via low bandwidth networks without sacrificing its comprehensibility. Video coding systems that use the particular structure of SL, i.e., the fact that the hands and face are the most important and thus need higher quality of representation, e.g., [1], [23]. More relevant to our work is [25], where a laboratory study was done to identify the lower threshold at which intelligible real-time conversations could be held. It was found that even at 5fps a discussion could be done, however the signers would have to sign at a slower speed. A limit of 10fps at 50kbps was found to be adequate for the general case.



(a) (b)

Figure 2: The consecutive frames from rest to rest position corresponding to the sign "καλημέρα" (good morning) (a) and the wrist trajectory superimposed in frame 384 (b)

The summarization of SL videos per se has been treated only in few works. In [15] the region of hands and face are segmented using skin color and then modeled using Zernike moments. The second derivative of the moment norm may be used to extract the keyframes, assuming these are the turning points in the overall motion. It gives reasonable results, however, the shapes are grossly represented, and the calculation of the moments may be demanding for better representation. Related to our method is [17], where the frames corresponding to the Maximum Curvature Points (MCPs) of the global trajectory is proposed to be taken as the keyframes for the compression of the sign's video and for solving the corresponding sign classification problem. However, this feature is invariant to the motion speed variability that may occur when different persons perform the same gesture making its use unsuitable for describing the dynamics of the motion model which are critical for selecting the keyframes of a gloss.

3 Problem Formulation

The inability of the human visual system to perceive and track fast motions has been extensively exploited to reduce the high computational cost in many applications in the field of computer graphics [4]. This is because the motion velocity, constituting a basic movement property, strongly affects the intelligibility of the action described by the motion. In a SL video there are some frames, which depict the hands clearly and others that suffer from motion blur. The blurry frames most often correspond to abrupt hand motion that the camera struggles to capture, but typically such frames are not semantically significant [11]. In addition, not all of the frames depicting the hands sharply are necessary for the intelligibility of the sign, since they may be just repeating the same (redundant) view. For example in Fig. 2 where the frames of sign "καλημέρα" (good morning) are depicted, the semantic meaning can be represented just by the frames 384 and 397, while the rest of them actually carry redundant information. This example is not exceptional, so in general it is reasonable to expect only a few semantically meaningful keyframes, while the majority of them are non critical at all. We are going to exploit that observation to extract summaries out of SL videos, without sacrificing their intelligibility.

In order to identify the keyframes of a SL video we are going to propose the t- parameterized curvature function of the trace of the signer's wrist. To this end, let us define the following pointset:

$$\mathbb{P}_N = \{\mathbf{r}_n\}_{n=1}^{N} \tag{1}$$

with $\mathbf{r}_n \in \mathbb{R}^2$ denoting the coordinates of the signer's wrist at the $n-th$ frame of the video

with respect to the frame's upper left corner. Note that the aforementioned pointset, denotes the trace of the signer's wrist in 2-D Euclidean space $\mathbb{R}^2$, as it moves from its start (rest) position to its end one as shown for the sign "καλημέρα" (good morning) in Fig.2 (b). In order to make mathematics tractable, let us consider that the elements of the pointset $\mathbb{P}_N$ result from sampling the following 2-D differentiable curve:

$$\mathbf{r}(t) = [x(t)\ y(t)]^T,\ t \in [t_0,\ T] \tag{2}$$

that represents the time - parameterized trajectory of a particle as it moves in the 2-D Euclidean space $\mathbb{R}^2$ for the time interval $[t_0,\ T]$. With $s(t)$ denoting the arc length that the particle has moved along the curve $\mathbf{r}(t)$ in the time interval $[t_0,\ t]$, that is:

$$s(t) = \int_{t_0}^{t} ||\mathbf{r}^{(1)}(\tau)||_2 d\tau, \tag{3}$$

with $||\mathbf{f}(t)||_2$ denoting the l_2 norm of 2-D vector function $\mathbf{f}(t)$, the 2-D curve defined in Eq. (2) can be reparameterized by its arc length s as:

$$\mathbf{r}(s) = [x(s)\ y(s)]^T,\ s \in [s(t_0),\ s(T)]. \tag{4}$$

Note that by parameterizing the trajectory of the particle by the arc length, its description does not depend on the rate quantified by the time derivative of the arc length, i.e. $s^{(1)}(t)$, in which the particle has traversed it. In other words, an infinity of t-parameterized trajectories correspond to the same s-parameterized one defined in Eq. (4).

3.1 Frenet-Serret Frame

Using the s-parameterized trajectory defined in Eq. (4), and using some differential geometry concepts, the Frenet-Serret **tnb** frame is defined through the following three orthogonal vectors:

$$\mathbf{t}(s) = \frac{\mathbf{r}^{(1)}(s)}{||\mathbf{r}^{(1)}(s)||_2}, \quad \mathbf{n}(s) = \frac{\mathbf{t}^{(1)}(s)}{||\mathbf{t}^{(1)}(s)||_2}, \quad \mathbf{b}(s) = \mathbf{t}(s) \times \mathbf{n}(s) \tag{5}$$

where $\mathbf{x}^{(1)}(s)$, "$\times$" denote the derivative with respect to arc length s of the 2-D curve $\mathbf{x}(s)$ and the cross product operator respectively. More precisely, the tangent vector $\mathbf{t}(s)$ points to the direction the curve travels, the normal vector $\mathbf{n}(s)$ is orthogonal to the tangent, while binormal vector $\mathbf{b}(s)$ is orthogonal to the plane defined by the aforementioned pair of orthogonal vectors, as it can be validated by Eq. (5). Note that the Frenet-Serret frame can be defined and used for the description of 2-D curves that are not straight lines where the curvature is equal to zero. Moreover, in the case of 2-D or plane curves the curvature at a point of such a differentiable curve is defined as the reciprocal of the radius of its osculating circle, that is the circle that best approximates the curve near that point. It is clear that the smaller this circle, the higher its curvature is, with its units being m^{-1}.

In order to take into account the existing particularities in the description of the 2-D curves and define an appropriate measure that uses in a physical manner the above mentioned quantities, using **tnb** frame, we can prove the following proposition.

Proposition 1: Let $\mathbf{r}^{(n)}(s),\ n = 1,\ 2,\ ||\mathbf{r}^{(1)}(t)||_2$ be the $n-$th order derivative of the s-parameterized trajectory of the curve defined in Eq. (4) with respect to its arc length s (i.e., the s-parameterized counterparts of the velocity and the acceleration respectively) and the speed $v(t)$ respectively. Then, the following s-parameterized based relation holds:

$$\mathbf{r}^{(1)}(s) \times \mathbf{r}^{(2)}(s) = \kappa(s)\mathbf{b}(s) \tag{6}$$

where "$\times$" denotes the cross product operator and $\kappa(s)$ is the curvature.

Using Proposition 1 and the unity norm of the vector $\mathbf{b}(s)$, we can define the absolute value of the curvature by the following relation:

$$|\kappa(s)| = ||\mathbf{r}^{(1)}(s) \times \mathbf{r}^{(2)}(s)||_2 \tag{7}$$

where $|x|$ denotes the absolute value of the scalar quantity x. Note that the units of the s-parameterized curvature $\kappa(s)$ are m^{-1} and as we can see from Eq. (7) defines the radius of an instantaneous circle where the motion should be done.

Let us now concentrate on the basic drawback of the above defined s- parameterized quantity. Notice that since the same displacement could result from the motion of the particle in the same time interval with a constant velocity, which is well known as **average velocity**, the s- parametrized curvature can be considered invariant to the particle's motion. This in turn means that s-parameterized curvature constitutes a proper **static** or **shape like** descriptor and it is unsuitable for describing the **kinematic model** of the particle. In order to make this last point clear, let us consider the following simple but informative example.

Circular motion: Let us consider that a particle moves on a circle of radius r with its motion model described by the phase function $\theta(t)$. In this case the trajectory of the particle is a 2-D curve and its s-parameterized form is defined by the following relation:

$$\mathbf{r}(s) = r \left[\; cos(\tfrac{s}{r}) \;\; sin(\tfrac{s}{r}) \; \right]^T$$

and the arc length $s(t)$, from Eq. (3), is given by $s(t) = \theta(t)$. By computing $\mathbf{r}^{(n)}(s)$, $n = 1, 2$ and using Eq. (7) we can find out that its curvature is given by the following relation:

$$\kappa(s) = r^{-1} \tag{8}$$

showing that the s-parameterized curvature $\kappa(s)$ does not depend on the motion model of the particle.

Moreover, we can prove a more general result regarding the s-parameterization based 2-D curves in the following proposition.

Proposition 2: Let $\mathbf{r}(t)$ be a t-parameterized trajectory of a particle that is moving into $\mathbb{R}^2$ for the time interval $[t_0, \; T]$ and $f(t)$ a continuous and increasing function of t in the same time interval with its initial and final value satisfying $f(t_0) = t_0$ and $f(T) = T$. Let us also consider that $\hat{\mathbf{r}}(t) = \mathbf{r}(f(t))$ be the t-parameterized trajectory of the particle resulting from the composition of the $\mathbf{r}(t)$ with function $f(\cdot)$. Then, their s-parameterized counterparts coincide, i.e., $\hat{\mathbf{r}}(s) = \mathbf{r}(s)$.

4 The Proposed Solution

We propose the use of the t-parameterized counterpart of the curvature. By using the chain rule, it is evident that it is defined as follows:

$$\kappa(t) = \kappa(s(t))v(t) \tag{9}$$

with $v(t)$ denoting the speed of the particle. It is clear that the units of the proposed re-parameterized curvature are sec^{-1}, i.e., Hz and consequently it can be considered as the instantaneous ordinary frequency of the motion of the particle as it moves on the 2-D space.

As we are going to see in the example of the simple circular motion, the proposed descriptor does not only represent the related **shape** information of the trajectory, but also the **kinetics**, that is the **dynamics**, of the motion.

Circular motion (continued): Let us now use the t-parameterized trajectory of the particle, i.e.:

$$\mathbf{r}(t) = r\left[\ cos(\theta(t))\ \ sin(\theta(t))\ \right]^T$$

Using Equs. (8), (9) and by computing the speed of the particle we can easily find out that the t-parameterized curvature $\kappa(t)$ is given by the following relation:

$$|\kappa(t)| = |\theta^{(1)}(t)|.$$

Note that the proposed t-parameterized curvature $\kappa(t)$ depends on the kinematic model of the particle and this in turn, justifies our proposition. Note also that, in the case of a circular motion with a constant angular velocity Ω_0, the proposed t parameterized curvature is equal to Ω_0.

It is clear that in the case of a 2-D trajectory, the video frames where the absolute value of the t-parameterized curvature $|\kappa(t)|$ attains its maxima, i.e.:

$$\kappa^{(1)}(t) = 0 \ \text{ and } \kappa^{(2)}(t) < 0 \tag{10}$$

is proposed to be the keyframes that can be used for summary extraction. We must stress at this point that the maxima of the s-parameterized curvature $|\kappa(s)|$ was proposed for identifying keyframes in [12].

Having defined all the necessary quantities, in the next section we are going to apply our figure of merit in real SL videos.

5 Experimental results

In this section we are going to present the results we have obtained from the application of the proposed figure of merit as well as other state of the art techniques [12, 15] in real Greek Sign Language Data.

5.1 Experimental Setup

To evaluate our method, we developed our own keyframe-annotated SL dataset, since to our knowledge there is no such available. It is composed by 32 videos of Greek SL of a total duration of 168 minutes containing approximately 5500 signs. Eight native signers performed four different scripts and captured from a Ximea camera with 60 fps. The scripts were realistic and concerned the scenario of interacting with a doctor via a video relay service using a vocabulary of 387 glosses. The videos contain continuous SL sentences that were manually annotated in gloss level by using the ELAN tool [8]. In addition, the dataset was annotated by four experts in Greek SL who selected the least amount of keyframes for each gloss in order to be fully understandable. The least amount of keyframes for each gloss had a median of two (2) and ranged between $[1,\ 10]$. The videos, along with their annotations are publicly available [1].

We used our criterion and $k(s)$ [12] in 2-D wrist's trajectories, using a similar setup. In addition both techniques were also compared to a Zernike moments-based technique [15]. As there is no standard protocol for the assessment of the extracted summaries we assessed the performance of the techniques by comparing them in (1) objective measures using the ground-truth, (2) human-based evaluation of understanding, and (3) gloss classification.

For the moments-based technique, that uses the whole video frames in its computations, we directly used the extrema of the curve $Z_p(t)$ instead of its second derivative proposed by

[1] https://vagsart.github.io/sl_keyframe_dataset/

the authors [15], because it performed better in our dataset. The pointset of the signer's wrist trajectory, that constitutes the first step of our overall framework (Fig. 1), was determined from MediaPipe skeleton tracker [16] and was smoothed by a Gaussian kernel to eliminate the noise influence occurred from the wrist detection algorithm. Afterwards, the Signing interval was estimated using the dominant hand position and the methods' curves (including Zernike moments based technique) were calculated between rest to rest intervals, which were identified from the relative position of the wrist to the manually selected signing area (Fig. 1). Subsequently, the informative points were identified, by applying the methods on the dataset.

Finally, in order to fairly compare the techniques we have defined the compression rate R_c as the desired number of the selected keyframes by each technique divided by the number of selected ones by the annotator. For every value of the compression rate R_c, the desired number of keyframes for each technique is selected based on the prominence value of their corresponding maxima. Note that since the strongest maxima of each method are attained in different time instances, the selected keyframes by each method are in general differently distributed along the glosses and this affects their performance.

5.2 Objective Measures Evaluation

We assessed the methods performance by comparing their selected keyframes with the ground truth one as we can see in Fig. 1, using the well-known *Recall* rate and F_2 score. In order to transform the problem at hand into a binary one, where we can use these metrics, we used the temporal distance between the keyframes and the ground truths. We consider that for each identified keyframe all neighbor frames that are in an absolute distance less than or equal to an experimentally defined threshold Δ are labeled by 1. Following such a procedure we obtain a binary label for every video frame we used for the evaluation. Because of the occurrence of many abrupt motions in SL, the threshold Δ in this experiment was set to 5 frames, which corresponds to 1/12 seconds since the camera frame rate was 60. For the impact of the threshold Δ to *Recall* rate and F_2 score, please see the supplementary material.

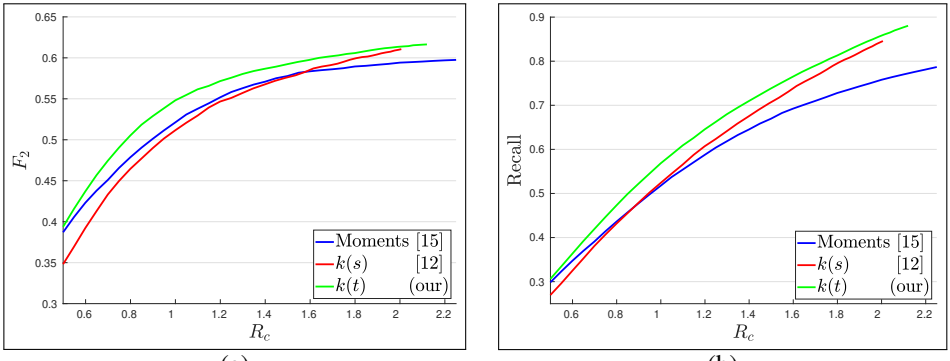
The obtained results, in terms of recall and the F_2 score versus the ratio R_c are shown in Fig. 3(a) and (b) respectively. It is evident that the proposed figure of merit outperformed the other methods. This in turn means that the proposed technique was in closer Δ proximity to the ground truth and thus captured the meaning more accurately. Indeed, when for example the ratio $R_c = 1$ the performance of the proposed t-parameterized curvature in terms of the F_2 score was 4% better than s-parameterized curvature and 3% better than the moment-based technique. By taking into account the complexity of the problem at hand, we must stress at this point that the achieved F_2 score by the proposed criterion is very promising.

Regarding the *Recall* rate, comparative results are depicted in Fig. 3(b). It is evident that the proposed technique achieved again better *Recall* rate than the other ones. Indeed, 57% of the selected, by the proposed technique, glosses were in Δ proximity with the corresponding of the annotators, while the s-parameterized curvature and moment-based technique 52%.

5.3 Human-Based Evaluation

In this experiment we evaluated the intelligibility of the extracted summaries by using human SL experts. We included the human-generated summaries as well, to evaluate subjectivity.

The evaluators were given only short videos depicting isolated signs and not the continuous sentences they were extracted from, because the latter are generally more easily interpretable due to their context. Context may lead to more concise summaries, but we left it for future work.



(a) (b)

Figure 3: Objective Measures Evaluation: Obtained results in terms of (a) F_2 score rate and (b) *Recall* versus R_c ratio for $\Delta = 5$

More specifically, the extracted summaries after they segmented in glosses (Fig. 1), were used for the reconstruction of the original videos by repeating every keyframe, until the occurrence of the next keyframe, so that the reconstructed video would have approximately the same duration with the original one. Those reconstructed videos, were evaluated by four expert interpreters in Greek SL who graded the depicted signs as understandable (2), non-understandable (0) or semi-understandable (1). Semi-understandable were those videos where the phonemic structure of the sign was not complete. In these videos a path movement or an internal (finger) movement was missing or an important handshape was missing. However, the lexical meaning was clear for the observer. The results of the human-based evaluation in approximately 500 signs of the dataset are shown in Table 1. The superiority of the proposed technique is evident, the 2-D t-parameterized curvature outperforms its s-parameterized counterpart thus validating our proposition.

5.4 Gloss Classification

In this last experiment we evaluated the above mentioned techniques as well as the extracted summaries by the human SL experts, that are considered as the ground truth keyframes, in the gloss classification problem. Specifically, given the keyframes of each technique for $R_c = 1$ and by exploiting the available annotation of glosses, that is the start and stop timestamps of every gloss along with its meaning, we can group the selected keyframes to glosses (Fig. 1) and then, the classification problem can be considered as a problem of identification of the meaning of the glosses contained in the available data.

To this end, four databases were created, one for each technique, containing the glosses with the features from their extracted keyframes. We used the keypoints from signer's both hands obtained from MediaPipe, as a feature of each gloss. Due to the difference in duration and the variability in the number of selected keyframes for each gloss, for the solution of the classification problem we used a GRU model [6]; it had 2 hidden layers of size 32, followed by a batch normalization and a linear layer. The model was implemented in PyTorch and trained for 10^4 epochs with SGD optimizer with learning rate 5×10^{-4} and a batch size of 256. The total parameters of the network were approximately 34×10^3.

By leaving out one (1) of the eight (8) signers (four videos), we divided the videos in train and test set. The total results was obtained by averaging for the eight splits. The results in terms of Top N accuracy for $N = 1, 2, 5$ and 10 are shown in Table 2. We consider the

	Techniques			
	Ground Truth	$k(s)$ [12]	Moments [15]	$k(t)$ (our)
Understandable	0.598	0.490	0.426	0.512
Semi-Understandable	0.254	0.238	0.272	0.254
Non-Understandable	0.148	0.272	0.302	0.234

Table 1: Human Based Evaluation: Proportion of glosses characterized from SL experts, as Understandable, Semi-Understandable and Non-Understandable from their keyframes

	Techniques			
	Ground Truth	$k(s)$ [12]	Moments [15]	$k(t)$ (our)
Top-1	0.56	0.39	0.38	**0.43**
Top-2	0.70	0.50	0.51	**0.54**
Top-5	0.82	0.62	0.64	**0.68**
Top-10	0.88	0.69	0.73	**0.75**

Table 2: Evaluation in classification task in the keyframe skeletal features obtained from proposed and techniques in [12, 15]

classification being Top - N accurate if the true meaning of the gloss belongs at least in the N most probable classes. It is evident that the keyframes of the proposed t-parameterized criterion are more suitable for identifying the gloss meaning as they are better in terms of the accuracy metric. The results are promising given the high complexity of the problem considering that the number of classes is 387.

5.5 Computational Complexity

Given a SL video i.e., a sequence of frames, the computational burden of the under comparison summarization schemes, heavily depends on the computational cost needed for constructing the 2-D wrist's trajectories, for the proposed and the spatial curvature [12] based one, and the 1-D decision sequence for the Zernike moments based technique [15]. Considering that those sequences are constructed in a frame by frame base, all the above mentioned techniques for identifying the keyframes demand basic operations, such as filtering and maxima detection whose computational cost is negligible. However, the computational costs needed for constructing the aforementioned sequences are totally different. Specifically, the construction of the t and s based trajectories can be achieved using MediaPipe, running in real-time on a CPU spending 6.8 msec per frame in our Intel(R) Core(TM) i7-7820X CPU @ 3.60GHz. On the other hand, the bottleneck in the technique proposed in [15] is the calculation of high-order Zernike moments that might be demanding in computational resources. Using our current implementation, Zernike moments calculation on a single frame can take upwards of 38.9 msec in the same configuration, which under the used frame rate does not permit the adoption of this technique in a real-time base.

6 Conclusion

In this work, we introduced a method for extracting keyframes from sign language videos, while preserving the intelligibility of the gestures. The time-parameterized Frennet-Serret frame for tracking the signer's wrist was used and the curvature of the trajectory, was proposed for the identification of the SL video keyframes. The applicability of the method was verified experimentally in sign language videos from a dataset containing 5000 signs and the better performance of the proposed technique against the s-parameterized curvature and a moments based technique was confirmed. The generalization of the proposed summarization scheme for the 3-D case does not constitute a trivial problem. In addition its combination with non-manual features that also convey information constitutes a critical issue. Both problems are currently under investigation.

7 Acknowledgments

This work was supported by the T1EΔK-01299 HealthSign project (www.healthsign.gr), which is implemented within the framework of "Competitiveness, Entrepreneurship and Innovation" (EPAnEK) Operational Programme 2014-2020, funded by the EU and national funds.
The authors would also like to thank Dr. Andrikopoulou Irini, MSc Stavropoulou Eugenia and PhD candidates Zacharopoulou Vasiliki and Mavrogiannaki Aresti, from Deaf Studies Unit of University of Patras, for their crucial participation in the human based evaluation experiment.

References

[1] Dimitris Agrafiotis, Nishan Canagarajah, David R. Bull, Jim Kyle, Helen Seers, and Matthew Dye. A perceptually optimised video coding system for sign language communication at low bit rates. *Signal Processing: Image Communication*, 21(7):531 – 549, 2006. ISSN 0923-5965. doi: https://doi.org/10.1016/j.image.2006.02.003. URL http://www.sciencedirect.com/science/article/pii/S0923596506000142.

[2] R. Agyeman, R. Muhammad, and G. S. Choi. Soccer video summarization using deep learning. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 270–273, March 2019. doi: 10.1109/MIPR.2019.00055.

[3] Z. ¯Cerneková, C. Nikou, and I. Pitas. Entropy metrics used for video summarization. In *Proceedings of the 18th Spring Conference on Computer Graphics*, SCCG '02, pages 73–82, New York, NY, USA, 2002. ACM. ISBN 1-58113-608-0. doi: 10.1145/584458.584471. URL http://doi.acm.org/10.1145/584458.584471.

[4] Alan Chalmers, Kirsten Cater, and David Maflioli. Visual attention models for producing high fidelity graphics efficiently. In *Proceedings of the 19th spring conference on Computer graphics*, pages 39–45, 2003.

[5] Vasileios Chasanis, Aristidis Likas, and Nikolaos Galatsanos. Efficient video shot summarization using an enhanced spectral clustering approach. In Véra Kůrková, Roman Neruda, and Jan Koutník, editors, *Artificial Neural Networks - ICANN 2008*, pages 847–856, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-87536-9.

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[7] O. Crasborn. Phonetic implementation of phonological categories in sign language of the netherlands. *Sign Language & Linguistics, Utrecht: Landelijke Onderzoeksschool Taalwetenschap*, 5(1), 2001. doi: 10.1075/sll.5.1.09cra.

[8] O. Crasborn and H. Sloetjes. Enhanced elan functionality for sign language corpora. In Sixth International, editor, *Proceedings of LREC 2008*. Conference on Language Resources and Evaluation, 2008.

[9] Brentari D. A prosodic model of sign language phonology. *Cambridge, MA: MIT Press*, 1998.

[10] Sandra Eliza Fontes De Avila, Ana Paula Brandao Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68, 2011.

[11] Mark Dilsizian, Polina Yanovich, Shu Wang, Carol Neidle, and Dimitris Metaxas. A new framework for sign language recognition based on 3D handshape identification and linguistic modeling. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1924–1929, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/1138_Paper.pdf.

[12] M Geetha and PV Aswathi. Dynamic gesture recognition of indian sign language considering local motion of hand using spatial location of key maximum curvature points. In *2013 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pages 86–91. IEEE, 2013.

[13] G. Guan, Z. Wang, K. Yu, S. Mei, M. He, and D. Feng. Video summarization with global and local features. In *2012 IEEE International Conference on Multimedia and Expo Workshops*, pages 570–575, July 2012. doi: 10.1109/ICMEW.2012.105.

[14] Dan Guo, Wengang Zhou, Anyang Li, Houqiang Li, and Meng Wang. Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation. *IEEE Transactions on Image Processing*, 29:1575–1590, 2019.

[15] D. I. Kosmopoulos, A. Doulamis, and N. Doulamis. Gesture-based video summarization. In *IEEE International Conference on Image Processing 2005*, volume 3, pages III–1220, Sep. 2005. doi: 10.1109/ICIP.2005.1530618.

[16] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019. URL http://arxiv.org/abs/1906.08172.

[17] B. Mahasseni, M. Lam, and S. Todorovic. Unsupervised video summarization with adversarial lstm networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2982–2991, July 2017. doi: 10.1109/CVPR.2017.318.

[18] Magda Nikolaraizi, Ioanna Vekiri, and Susan R. Easterbrooks. Investigating deaf students use of visual multimedia resources in reading comprehension. *American Annals of the Deaf*, 157(5):458–473, 2013. doi: 10.1353/aad.2013.0007.

[19] R. Pfau, M. Steinbach, and B. Woll. Sign language: An international handbook. *Berlin: De Gruyter Mouton*, 2012. doi: 10.1075/sll.5.1.09cra.

[20] Po Kong Lai, M. Décombas, K. Moutet, and R. Laganière. Video summarization of surveillance cameras. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 286–294, Aug 2016. doi: 10.1109/AVSS.2016.7738018.

[21] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *European conference on computer vision*, pages 540–555. Springer, 2014.

[22] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. *ArXiv*, abs/1805.10538, 2018.

[23] D. M. Saxe and R. A. Foulds. Robust region of interest coding for improved sign language telecommunication. *IEEE Transactions on Information Technology in Biomedicine*, 6(4):310–316, Dec 2002. doi: 10.1109/TITB.2002.806094.

[24] L. Shao and L. Ji. Motion histogram analysis based key frame extraction for human action/activity representation. In *2009 Canadian Conference on Computer and Robot Vision*, pages 88–92, May 2009. doi: 10.1109/CRV.2009.36.

[25] Jessica J. Tran, Ben Flowers, Eve A. Risken, Richard E. Ladner, and Jacob O. Wobbrock. Analyzing the intelligibility of real-time mobile sign language video transmitted below recommended standards. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '14, pages 177–184, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2720-6. doi: 10.1145/2661334.2661358. URL http://doi.acm.org/10.1145/2661334.2661358.

[26] Huan Yang, Baoyuan Wang, Stephen Lin, David Wipf, Minyi Guo, and Baining Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4633–4641, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.526. URL http://dx.doi.org/10.1109/ICCV.2015.526.

[27] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 982–990, 2016.

[28] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV*, 2016.