Supplementary Material: A 2-D Wrist Motion Based Sign Language Video Summarization

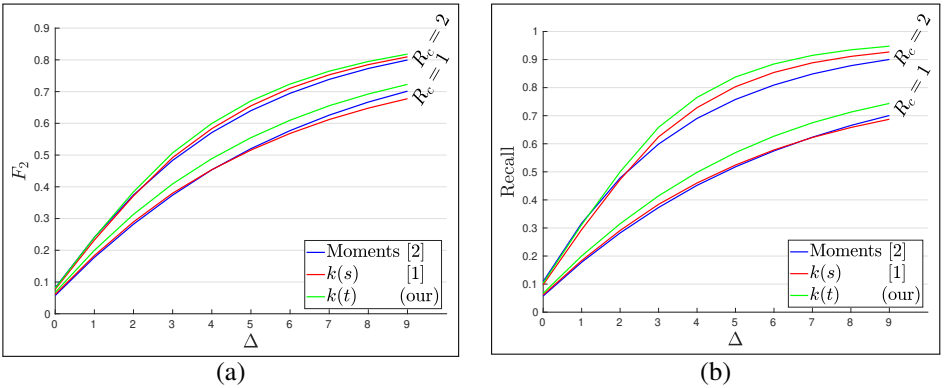
Evangelos G. Sartinas[1]
sartinas@ceid.upatras.gr
Emmanouil Z. Psarakis[1]
psarakis@ceid.upatras.gr
Klimis Antzakas[2]
k.antzakas@upatras.gr
Dimitrios I. Kosmopoulos[1]
dkosmo@upatras.gr

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

A Additional Experiments

A.1 The Impact of Threshold Δ on the Objective Evaluation Metrics

In order to quantify the impact of threshold Δ we measured for its different values the F_2 score and *Recall* rates for $R_c = 1$ and $R_c = 2$ (i.e., the amount of keyframes was set to be equal or twice as much as the ground truth one). The results are shown in Fig. 1. Clearly, both F_2 score and *Recall* rate, are increasing functions of the proximity factor Δ, retaining methods ordering as well. In addition, as it was expected both metrics are increasing as the ratio R_c increases.



(a) (b)

Figure 1: Obtained results in terms of (a) F_2 score rate and (b) *Recall* versus temporal proximity threshold Δ for $R_c = 1, 2$

		Techniques			
		Ground Truth	$k(s)$ [1]	Moments [2]	$k(t)$ (our)
Dominant Hand	Top-1	0.52	0.36	0.35	**0.39**
	Top-2	0.66	0.47	0.46	**0.50**
	Top-5	0.79	0.59	0.60	**0.64**
	Top-10	0.86	0.66	0.69	**0.72**
Dominant Hand + Pose	Top-1	0.54	0.37	0.37	**0.41**
	Top-2	0.68	0.48	0.49	**0.53**
	Top-5	0.80	0.60	0.62	**0.66**
	Top-10	0.87	0.67	0.70	**0.73**
Both Hands	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 1: Evaluation in classification task in the keyframe dominant manual (rows 1-4), dominant manual and non-manual (rows 5-8) and both manual (rows 9-12) skeletal features obtained from the Ground Truth, the Proposed and Techniques presented in [1, 2]

Note also that for both values of R_c and for both objective evaluation metrics the proposed technique outperforms the other ones. The superiority of the proposed technique is more evident in the *Recall* rate metric, that can be viewed as the probability that a relevant keyframe is selected by the technique and whose significance over *Precision*, that can be considered as the probability that a keyframe randomly selected from the pool of total selected keyframes is relevant, has been indicated from the use of F_2 score instead of F_1 in our evaluation.

A.2 GRU based Gloss Classification

In this experiment we evaluated the under comparison 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 in using different manual and non-manual skeletal features. We follow the framework that we describe in Subsection 5.4 of the main manuscript and use the same GRU model-based classification scheme. 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, the classification problem can be considered as a problem of identification of the meaning of the glosses contained in the available data.

The GRU model has the same configuration described in the manuscript and differs only in the input's features. Specifically, we have considered the following three (3) different sets of features for feeding the GRU based classifier:

- *Dominant manual features*: The cardinality of this set of features is $N_f = 21$ and contains all the dominant hand 3D keypoints identified by the skeleton tracker in each keyframe of the encoded gloss.

- *Dominant manual and non-manual features*: The cardinality of this set of features is $N_f = 46\ (21+25)$ and contains all the dominant hand as well as the pose 3D keypoints identified by the skeleton tracker in each keyframe of the encoded gloss.

- *Both manual features*: For this set of features the deep architecture of the GRU based classification neural net has been presented in Subsection 5.4 of the manuscript, but it is also included here for readability purposes.

The results in terms of Top N accuracy for $N = 1,2,5$ and 10 are shown in Table 1. We consider the 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. Note also that as the cardinality of the feature set increases, the performances of the under comparison techniques are also improved.

A.3 k-NN Based Gloss Classification

In addition to the GRU model based gloss classifier we also tested a simpler k-NN classifier. The goal of this experiment is to evaluate the performance of the summarization schemes in the gloss classification problem using a non-parametric classifier. To this end, in order to compensate the different duration as well as the variability in the number of selected keyframes for each gloss, we used dynamic time warping (DTW) [3] for measuring the distance between two encoded glosses, that is:

$$\mathcal{G}_i = \{K_{i,j}\}, j \in \mathcal{S}_i = \{1,2,\cdots,N_i\},\ i = 1,2$$

with N_1, N_2 keyframes respectively and with each $K_{i,j}$ being a $3 \times N_f$ matrix containing the 3D coordinates of the corresponding N_f dominant hand keypoints to the specific j-th keyframe of the G_i gloss.

In order to define the necessary by DTW local distance measure $d(K_{1,j}, K_{2,l})$ for the feature matrices $K_{1,j}, K_{2,l}, j \in \mathcal{S}_1,\ l \in \mathcal{S}_2$ we can solve the following well known Orthogonal Procrustes [4] optimization problem:

$$\begin{aligned} R^\star \quad = \quad & arg\min_R ||K_{1,j} - RK_{2,l}||_F \\ & \text{subject to:} \\ & R^T R = I \end{aligned} \tag{1}$$

with $||X||_F$ denoting the *Frobenius* norm of matrix X. Note thought that the same handshape, but with substantially different orientation, can be used in different glosses as we can see in the example shown in Figure 2. This in turn means that we can erroneously align totally different glosses in meaning. Thus, we define the required measure as follow:

$$d(K_{1,j}, K_{2,l}) = \begin{cases} ||K_{1,j} - R^\star K_{2,l}||_F, & if\ |\theta^\star| \leq \theta_{\max} \\ ||K_{1,j} - K_{2,l}||_F, & otherwise. \end{cases} \tag{2}$$

That is, for avoiding mismatches of the above mentioned kind we use the optimal solution of the Problem (1) in the definition of the local measure, only if the optimal rotation angle

(a) (b)

Figure 2: Two signs with the same handshape but different hand orientation and meaning, (a) number "one" and (b) personal pronounces "she"/"he"

$\theta^\star$ given by [5]:

$$\theta^\star = \frac{180}{\pi} \arccos\left(\frac{\text{trace}\{R^\star\} - 1}{2}\right) \tag{3}$$

is small.

In all experiments we have conducted the $\theta_{\max}$ was set to 5° while the best value of the hyperparameter k was found after grid searching and set equal to 5. The Top N results for $N = 1, 2, 5$ and 10 we obtained from the application of the k-NN based classifier for the under comparison techniques, are documented in Table 2. It is clear from the contents of this table that the keyframes of the proposed t-parameterized criterion are more suitable for identifying the gloss meaning even using such a simple classification method. We must stress at this point that although the obtained results by this simple k-NN schemes in some sense could be considered comparable with that obtained by the first classification scheme that was based on the GRU model, it can not be easily extended in using more skeletal features such as the keypoints of the other hand as well as the pose are. This is because such an extension demands the appropriate weighting of the manual and non-manual keypoints and that constitutes a non trivial task. Finally, note the inferior achieved performances by all the techniques in the case of the non-parametric k-NN based classifier as they are compared to the corresponding ones obtained by the GRU based classifier.

	Techniques			
	Ground Truth	$k(s)$ [1]	Moments [2]	$k(t)$ (our)
Top-1	0.49	0.32	0.30	**0.35**
Top-2	0.62	0.43	0.41	**0.46**
Top-5	0.75	0.54	0.54	**0.59**
Top-10	0.82	0.63	0.63	**0.67**

Table 2: Evaluation in classification task using a k-NN ($k = 5$) classifier in the keyframe dominant manual skeletal features obtained from the Ground Truth, the Proposed and Techniques presented in [1, 2]

References

[1] 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.

[2] 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.

[3] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[4] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[5] David J. Taylor, Camillo J.; Kriegman. Minimization on the lie group so(3) and related manifolds. In *Technical Report No. 9405*. Yale University, 1994.