AUTOMATIC ALIGNMENT BETWEEN SIGN LANGUAGE VIDEOS AND MOTION CAPTURE DATA: A MOTION ENERGY-BASED APPROACH

Fabrizio Nunnari, Mina Ameli, Shailesh Mishra

German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus D3 2

ABSTRACT

In this paper, we propose a method for the automatic alignment of sign language videos and their corresponding motion capture data, useful for the preparation of multi-modal sign language corpora. First, we extract an estimate of the motion energy from both the video and the motion capture data. Second, we align the two curves to minimize their distance. Our tests show that it is possible to achieve a mean absolute error as low as 1.11 frames using optical flow for video energy extraction and a set of 22 bones for skeletal energy extraction.

***Index Terms*—** sign language, motion capture, alignment, optical flow, motion energy, synchronization.

1. INTRODUCTION

Alignment of different modalities plays a significant role in many domains like autonomous driving [1], speech recognition [2], curve matching [3], sign language processing [4], and action recognition [5].

This paper addresses the problem of aligning two motion data sources that were captured simultaneously, on the same subject, but without a dedicated hardware synchronization mechanism: i) the frontal video of the subject performing sign language utterances, and ii) his/her skeletal animation recorded through an sensor-based (optical or inertial) full-body motion capture (MoCap) suit. For some applications, like extracting facial animation data from the video and merging them with the body motion to animate an avatar, an alignment between video and MoCap data is needed at frame-level precision. However, the manual post-processing of the two sources may be too time-consuming.

To overcome this challenge, we propose a novel alignment algorithm based on *motion energy* analysis. The main idea is to "summarize" the quantity of motion present in each data source into a single mono-dimensional signal, and then estimate the offset between the two modalities by shifting the two curves to find a good overlap. The motion energy of the MoCap data is calculated by an analysis of the movement of the bones, while the motion energy of the videos is computed through *frame differencing* or *optical flow* filtering.

Work funded by the German Federal Ministry of Education and Research (BMBF) through AVASAG (16SV8491) and BIGEKO (16SV9093) projects.

2. RELATED WORK

Previous studies on sequence alignments typically estimate the camera's underlying geometry directly or indirectly [6]. Multi-camera setups can use 2D homography-based geometry constraints to align multiple sequences, which have shown effectiveness in [7, 8]. Additionally, [9] introduced a new technique that aligns two sequences by exclusively utilizing the motion signals of matching Epipolar lines. However, these methods have limitations since they heavily rely on camera geometry not suitable for motion capture or video alignment.

Several techniques have been proposed for alignment in various applications. For example, a low-dimensional embedding using a weighted PCA algorithm was identified in [10], but it has limitations in dealing with multi-modal and noisy data. Neural networks have also been used for video synchronization [11], for instance in [12], which reformulates the problem as classification, but it requires a large amount of labeled training data.

Another strategy is to combine dynamic time warping (DTW) with certain regression models [13]. Junejo et al. [5] proposed a view-independent descriptor for video alignment using DTW while Zhou et al. [14] suggested the use of the Generalized Time Wrapping (GTW) approach, which can handle different modalities. Additionally, Chen et al. [15] aligned skeleton sequences from a Kinect RGB-D and a motion capture system using feature extraction and sub-sequence DTW, but they do not report an analysis between the ideal frame offset and the one determined by their algorithm. Despite the benefits of these approaches, their effectiveness in alignment is dependent on the choice of features.

To overcome the limitations of prior studies, we investigate motion energy-based alignment between video and motion capture data. Motion Energy Analysis (MEA) is a method commonly employed in the fields of social signal analysis and psychology to gauge human activities [16, 17]. MEA is based on calculating the difference between consecutive video frames to estimate the degree of motion of a human subject within a given frame [18]. This approach has several advantages, including its simplicity, as it only requires a fixed camera to prevent the motion of the subject from being mistaken for background movement, and its effectiveness in handling complex and varied motions in video data.

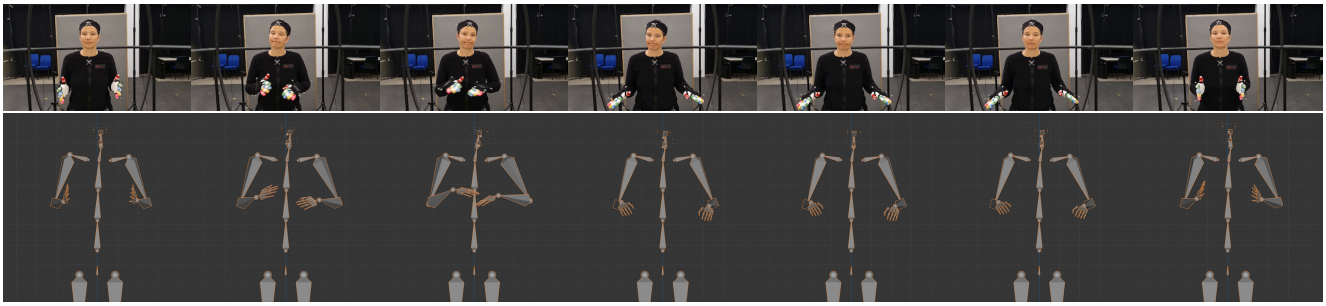

Fig. 1. Frame sequence extracted from sign NICHT, including the preparation and release phases. Top: the RGB video. Bottom: skeletal animation as seen from the Blender 3D editor.

3. METHOD

Data: Video Clips. To test our approach, we use 27 video clips showing insulated signs, for a total of 2415 frames, selected via random selection from an in-construction corpus of German sign language (DGS). Figure 1, top, shows an example. As a prerequisite, it is important that video captures are performed with a camera fixed in space, to avoid recognizing motion energy from the moving background. In general, gesture execution can be segmented into three phases: preparation, stroke, and release. Each video clip has been manually annotated for the beginning and end of the *stroke* phase. This has been done by considering the motion of the hands and any cues of activation of the body muscles.

Motion Energy from Video: Frame Differencing. Frame differencing compares consecutive video frames and quantifies the differences between them [18]. This method does not capture details about the direction or form of movement: it only measures the degree of change over time [17]. This is performed by computing the sum of the absolute differences between consecutive frames [19]. Given a frame i and the previous frame j, the motion `energy` for i is computed with the following Python code using 3D numpy arrays and the OpenCV framework [20]:

```
# Computes the difference in RGB space
diff = cv2.absdiff(frame_j, frame_i)
# Convert the difference in Grey space
img_gray = cv2.cvtColor(diff, cv2.COLOR_BGR2GRAY)
# Gets the median in a 5x5 area around the pixel.
filt = scipy.ndimage.median_filter(img_gray, size=5)
# Convert into a binary image with pixels at 0 or 255.
_, thresh = cv2.threshold(src=filt, thresh=15,
                  maxval=255, type=cv2.THRESH_BINARY)
# Sum up the "white pixels"
energy = thresh.sum()
```

Motion Energy from Video: Optical Flow. The Gunnar Farneback's algorithm for dense optical flow estimation computes the flow vectors for each pixel in the image by comparing the intensities of corresponding pixels in two consecutive frames [21]. The algorithm has been shown to be effective in a variety of computer vision applications, such as object tracking and motion estimation [22]. In our experiments, the motion `energy` from the optical flow is computed by the following Python code:

```
# Compute dense optical flow between two frames
gray_i = cv2.cvtColor(frame_i, cv2.COLOR_BGR2GRAY)
gray_j = cv2.cvtColor(frame_j, cv2.COLOR_BGR2GRAY)
flow = cv2.calcOpticalFlowFarneback(gray_i, gray_j,
                        None, 0.5, 3, 15, 3, 5, 1.2, 0)
# Compute the flow vectors
magnitude, angle = \
     cv2.cartToPolar(flow[..., 0], flow[..., 1])
# Sum up the magnitudes
energy = magnitude.sum()
```

Data: Skeletal Motion Capture. For each sign video clip, a corresponding "skeletal motion" clip is saved as BVH file (Biovision Hierarchy). Figure 1, bottom, shows an example. In digital computer graphics, a human skeleton (or better, an approximation of it) is a hierarchical tree structure describing how bones are connected to each other. The skeleton used in our corpus contains 98 bones, of which 51 control the upper body, head, and hands (including fingers). Each bone is defined by a *head* and a *tail*. Rotating the head of a bone affects the position in space of its tail and of the children's bones. For the testing purposes of this work, all of the 27 motion clips have been annotated with their correct frame offset (ground truth) with respect to the corresponding video. Importantly, for this setup, we know that the offset between the video and the motion clip lies between +/- 10 frames.

Motion Energy from Skeletal Motion Capture. Extracting the "motion energy" of a moving skeleton can be interpreted in different arbitrary ways. Our goal was to identify a strategy that leads to a correspondence between the motion extracted from the skeleton and the motion measurable from the corresponding video stream. Measuring the rotation of bones would result in a big discrepancy with the number of pixels changed in the video frames. For example, a quick

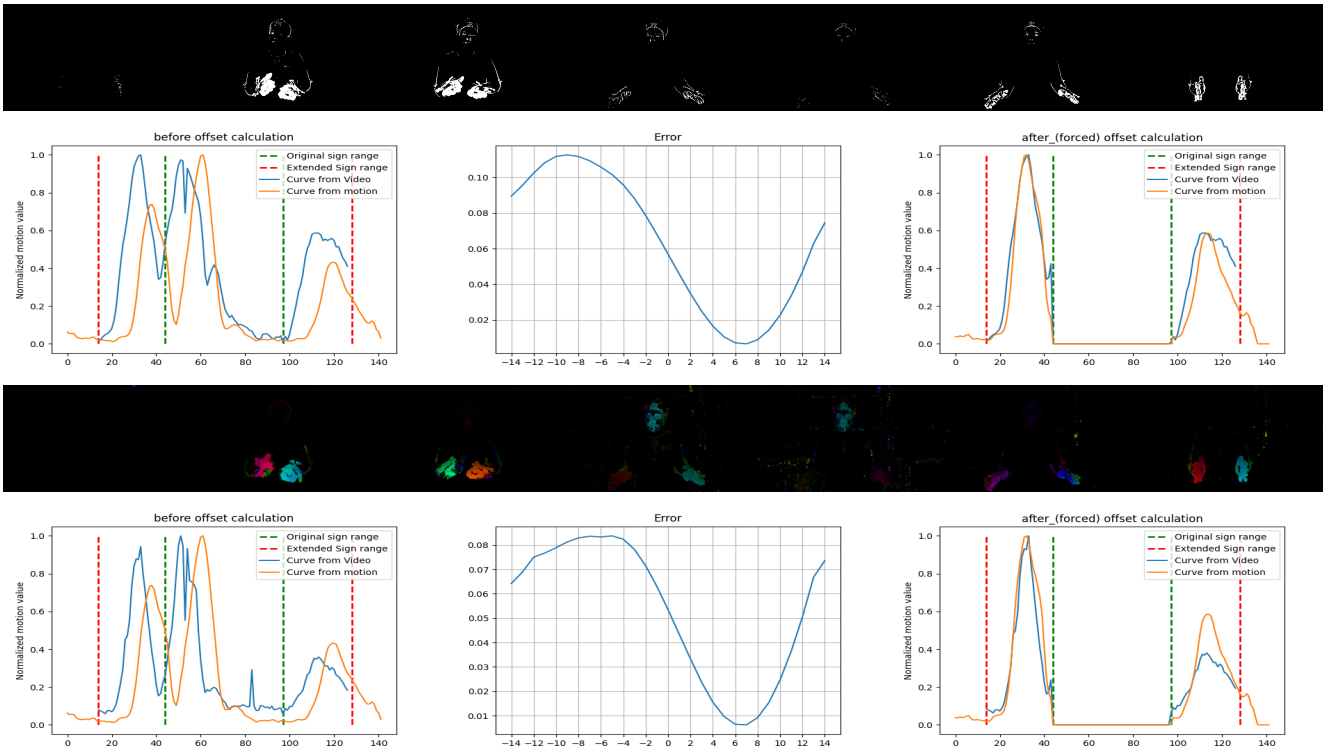

Fig. 2. Results for sign NICHT with FrameDiff (top) and OptFlow (bottom) filters. The images show frames highlighting motion energy activation areas. Left plots show video and skeletal motion energy curves, highlighting the internal *stroke range* and the extended *video padding*. Middle plots show the curves difference as a function of the offset applied to the video curve. Right plots show the curves after applying the best offset, forcing the *stroke* range to 0, and re-normalizing.

movement of the fingers would correspond to a lot of motion energy that reflects in a small number of pixel changes. Rather, we extracted motion energy by measuring at each frame, *speed of bone tails*. The speed of all tails is then summed together and used as body motion energy estimation.

Curves comparison. To estimate the correct offset between the video and motion energy curves, the two curves are shifted and the offset that minimizes their difference is selected. This difference is calculated as the mean of the difference between the two curves at each frame. However, since the techniques used to compute the two motion energy curves are not comparable, it is crucial to normalize the curves to a common amplitude range before computing the difference.

4. EXPERIMENTS

We conducted a series of experiments aimed at finding the best combination of parameters able to minimize the error between the correct offset and its estimation. We combined the following variables: **Filter**, motion estimation technique between *FrameDiff* and *OptFlow*; **Video and Motion Padding**, number of frames before and after the stroke range, controlling how much to look in the preparation and release phases; **Video visibility**, between *FULL*, if the full video was analyzed, and *RIGHT_ARM_TORSO*, when only the dominant arm and part of the torso are visible; **Bone set** indicates the list of bones that were used to compute the skeletal motion energy; **Motion Phase**, between *both*, *preparation*, and *release*, indicating which phase(s) were used to compute the difference between the two motion energy curves.

The experimental procedure is the following:

Step 1: compute the video motion energy curve using one of the two filtering methods. It is computed on the sign stroke range extended by the *Video Padding*.

Step 2: compute the skeletal motion energy curve using a given set of bones. It is computed on the stroke range extended by the *Motion Padding*;

Step 3: find the optimal offset by shifting the video curve inside the time range of the motion curve with steps of 1 frame (41 positions). For each tested offset: **i)** Flatten to 0 the stroke time range; **ii)** Re-normalize the curves in amplitude range $[0, 1]$; **iii)** Compute the difference between the curves.

Figure 2 visualizes an example of the experimental procedure for one sign. It shows a common pattern in the profile of the curves: a high peak of energy during the preparation

Table 1. Offset estimation error resulting from our experiments (the lower the better).

				Error (frames)			
Video filter	Padding (Video, Motion)	Video visibility	Bone set	Min	Max	MAE	MSE
FrameDiff	20, 40	FULL	HANDS_HEAD_5	0	11	2,59	12,96
	30, 50	FULL	HANDS_HEAD_5	0	7	1,74	6,26
	30, 50	FULL	RIGHT_ARM_3	0	7	1,93	7,04
	30, 50	FULL	HANDS_ELBOWS_HEAD_7	0	8	1,96	7,30
	30, 50	FULL	UPPER_BODY_22	0	5	**1,44**	**3,96**
	30, 50	RIGHT_ARM_TORSO	HANDS_HEAD_5	0	9	1,89	7,30
	30, 50	RIGHT_ARM_TORSO	RIGHT_ARM_3	0	8	1,74	6,48
	30, 50	RIGHT_ARM_TORSO	RIGHT_ARM_TORSO_7	0	8	1,74	6,41
OptFlow	30, 50	FULL	UPPER_BODY_22	0	5	**1,11**	**2,67**
	30, 50	FULL	HANDS_ELBOWS_HEAD_7	0	6	1,56	4,74
	30, 50	RIGHT_ARM_TORSO	RIGHT_ARM_3	0	7	1,48	5,11
	30, 50	RIGHT_ARM_TORSO	RIGHT_ARM_TORSO_7	0	7	1,52	5,07

phase (when the hands move from the rest position to the sign start), a second peak with the energy of the sign stroke, and a final peak when the hands release back to the rest position.

Two points deserve attention. First, after some initial observation of the motion curve graphs, we realized that the stroke range should be eliminated for our tests, mainly because the motion of the face and lips was generating a quantity of motion energy that can not be inferred from the skeletal motion data. Hence, we choose to *rely solely on the preparation and release phases*. Second, the re-normalization of the curves after removing the central *stroke range* is needed because we noticed that, in some cases, the stroke of a sign gives spikes of energy that make the two curves operate on two different scales.

Table 1 lists the results for all the tested combinations of parameters. For brevity, results are reported only for Motion Phase *both*, but in general, we measured that: i) using only one of the two phases (*preparation* or *release* worsens the results, and that ii) using only of the *preparation* phase gives always better results with respect to using only the motion energy of the *release* phase. A first test with *FrameDiff* filter, 5 bones, and video padding 20 was immediately improved by extending the video padding to 30, thus capturing full preparation and release phases in 30 frames (0.5 seconds). Further tests, removing or adding bones, lead to the best result. using 22 bones (arms, head, torso, and fingertips), with 1.44 MAE (mean absolute error) and 3.96 MSE (mean squared error). Other experiments, hiding part of the video to show only the dominant arm and torso, did not give improvements. Results were further improved by using the *OptFlow* filter, leading to **MAE 1.11** and **MSE 2.67**. This worsen to MAE 1.67 and MSE 6.93 (not in the table) when using only the preparation phase. Considering the desired MAE being 0, and that even for human annotators an error of 1 frame is often possible, our results can be considered satisfactory.

All the tests were conducted on a personal laptop with an 8-core i9 CPU and 32GB Ram. While the extraction of the motion energy from the skeletal data takes less than a couple of seconds, the computation of the motion energy from the video is more demanding. We compared the time needed to compute the motion energy for all of the 27 signs in our test set. OptFlow takes 1108 seconds (2.18 FPS), while FrameDiff takes 703 seconds (3.44 FPS); 58% faster than OptFlow.

5. CONCLUSION

We presented a method to automatically align video and motion capture data applied to the domain of sign language corpus creation. The method is based on the application of video/motion filters computing 1D curves "summarizing" the energy motion of both data sources and minimizing the distance between such curves. As such, the approach doesn't require any training data. To overcome the inconsistencies between the video facial motion and the skeletal motion, we realized that analyzing only the preparation phase of a sign execution is enough to achieve remarkable results. From our tests, optical flow is slower but more accurate than the frame differencing approach.

The main limitation of this approach is the requirement of a fixed camera, to avoid generating motion from the background. The method was tested by assuming the presence of time markers for the preparation and release phases. However, such markers would not be needed if the video motion energy of the face would be somehow measurable in the motion energy of the motion capture data.

Also, in our test procedure, we knew in advance the maximum difference between video and motion, allowing us to use a "brute-force" approach and identify the minimum error by testing on the whole range of possible offsets (41 tests in total). When such an estimate is not known, and the range of offset is potentially bigger, it might be more convenient to employ a general-purpose minimization algorithm, of course, at the cost of risking the identification of local minima.

To our knowledge, this is the first work employing a motion energy analysis in the context of sign language. This achievement in the problem of frame alignment suggests that this measure could help when performing other sign language motion-related tasks, such as sign spotting or sentence segmentation.

6. REFERENCES

[1] Ramzi Abou-Jaoude, "Acc radar sensor technology, test requirements, and test solutions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 4, no. 3, pp. 115–122, 2003.

[2] Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of speech recognition*, Prentice-Hall, Inc., 1993.

[3] Thomas B. Sebastian, Philip N. Klein, and Benjamin B. Kimia, "On aligning curves," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 1, pp. 116–125, 2003.

[4] Ulrich Von Agris, Jörg Zieren, Ulrich Canzler, Britta Bauer, and Karl-Friedrich Kraiss, "Recent developments in visual sign language recognition," *Universal Access in the Information Society*, vol. 6, pp. 323–362, 2008.

[5] Imran N Junejo, Emilie Dexter, Ivan Laptev, and Patrick Perez, "View-independent action recognition from temporal self-similarities," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 1, pp. 172–185, 2010.

[6] Cen Rao, Alexei Gritai, Mubarak Shah, and Tanveer Syeda-Mahmood, "View-invariant alignment and matching of video sequences," in *Computer Vision, IEEE International Conference on*. IEEE Computer Society, 2003, vol. 3, pp. 939–939.

[7] Flavio Padua, Rodrigo Carceroni, Geraldo Santos, and Kiriakos Kutulakos, "Linear sequence-to-sequence alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 304–320, 2008.

[8] Yaron Caspi and Michal Irani, "Spatio-temporal alignment of sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 11, pp. 1409–1424, 2002.

[9] Dmitry Pundik and Yael Moses, "Video synchronization using temporal signals from epipolar lines," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III 11*. Springer, 2010, pp. 15–28.

[10] Alexis Heloir, Nicolas Courty, Sylvie Gibet, and Franck Multon, "Temporal alignment of communicative gesture sequences," *Computer animation and virtual worlds*, vol. 17, no. 3-4, pp. 347–357, 2006.

[11] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu, "Tdan: Temporally-deformable alignment network for video super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3360–3369.

[12] Xinyi Wu, Zhenyao Wu, Yujun Zhang, Lili Ju, and Song Wang, "Multi-video temporal synchronization by matching pose features of shared moving subjects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[13] Eugene Hsu, Kari Pulli, and Jovan Popović, "Style translation for human motion," in *ACM SIGGRAPH 2005 Papers*, pp. 1082–1089. 2005.

[14] Feng Zhou and Fernando De la Torre, "Generalized time warping for multi-modal alignment of human motion," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1282–1289.

[15] Xi Chen and Markus Koskela, "Sequence alignment for rgb-d and motion capture skeletons," in *Image Analysis and Recognition: 10th International Conference, ICIAR 2013, Póvoa do Varzim, Portugal, June 26-28, 2013. Proceedings 10*. Springer, 2013, pp. 630–639.

[16] Fabian Ramseyer and Wolfgang Tschacher, "Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome.," *Journal of Consulting and Clinical Psychology*, vol. 79, no. 3, pp. 284–295, 2011.

[17] Fabian T Ramseyer, "Motion energy analysis (mea): A primer on the assessment of motion from video.," *Journal of counseling psychology*, vol. 67, no. 4, pp. 536, 2020.

[18] Milan Sonka, Vaclav Hlavac, and Roger Boyle, *Image processing, analysis, and machine vision*, Cengage Learning, 2014.

[19] Karl Grammer, Kirsten Kruck, Astrid Juette, and Bernhard Fink, "Non-verbal behavior as courtship signals: The role of control and choice in selecting partners," *Evolution and Human Behavior*, vol. 21, no. 6, pp. 371–390, 2000.

[20] Gary Bradski, Adrian Kaehler, et al., "Opencv," *Dr. Dobb's journal of software tools*, vol. 3, no. 2, 2000.

[21] Gunnar Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*. Springer, 2003, pp. 363–370.

[22] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski, "A database and evaluation methodology for optical flow," *International journal of computer vision*, vol. 92, pp. 1–31, 2011.