Natural Gesture Extraction Based on Hand Trajectory

Naoto Ienaga[1], Bryan W. Scotney[2], Hideo Saito[1], Alice Cravotta[3], and M. Grazia Busà[3]

[1]*Graduate School of Science and Technology, Keio University, Japan*
[2]*School of Computing, Ulster University, Northern Ireland*
[3]*Department of Linguistic and Literary Studies, Padova University, Italy*

Abstract

Automatic natural gesture recognition can be useful both for the development of human-robot applications and as an aid in the study of human gesture. The goal of this study is to recognize natural gestures using only an RGB video without machine learning methods. To develop and test the proposed method we recorded videos in which a speaker gestured naturally but in a controlled way. The advantage of using this method over lab-recorded data is that the data contain variations in gestures that are typically encountered when analyzing gestures of TV news or speech videos on the Internet. The hand positions are computed by a pose estimation method, and we recognize the gestures based on the hand trajectories, assuming that the gesturing hand(s) do(es) not change its direction abruptly during each phase of a gesture. Based on ground-truth annotations provided by linguistic experts, the accuracies were 92.15%, 91.76% and 75.81% for three natural gestures selected.

Keywords: Gesture recognition, video annotation

1 Introduction

In scientific fields where natural gestures are studied as part of human cognitive, linguistic and communicative processes, gestures are annotated manually in a time-consuming manner. Automatic gesture recognition can contribute to reducing annotation time. Natural gesture recognition also has an advantage over lab-recorded data since it can be used for recognizing speakers' gestures in TV news or speech videos on the Internet. In this paper, we report the results of a study aimed at detecting natural gestures in discourse, using natural speech recorded in a controlled setting.

Gestures can be intended as hand movements that accompany speech in human communication. In gesture studies, every hand gesture has been seen as a sequence of different phases [Kendon, 1980]. The gesture starts from the moment in which the arm begins to depart from a position of relaxation (rest position) until the moment when it returns to one rest position (retraction). The peak of the hand movement is called stroke. It is the peak of maximum effort [Dell, 1970], the moment when the movement dynamics, shape and meaning are manifested with greatest clarity. The gesture stroke usually aligns with the pronunciation of the word or sentence that constitutes the semantic nucleus of the speech. Many strokes can be performed before the hand comes back to a rest position. Gestures change their meaning and pragmatic functions depending on the hand shapes, palm orientation, trajectory/movement, velocity etc., with a certain degree of variability due to cultural and individual differences. Gestures can be are categorized along a *continuum* [McNeill, 2008] depending on different dimensions: (1) the extent to which they are conventionalized (convention means that the pair of gestures and meanings meet some kind of socially constituted or collective standard); (2) the presence/absence of accompanying speech; (3) relationship to linguistic properties (e.g., constraints due to language properties).

On one pole of this continuum [McNeill, 2008] places signs (used in sign language) that are fully conventionalized, have linguistic properties (are actually part of a language), and never occur with speech. Natural

 ISBN 978-0-9934207-3-3

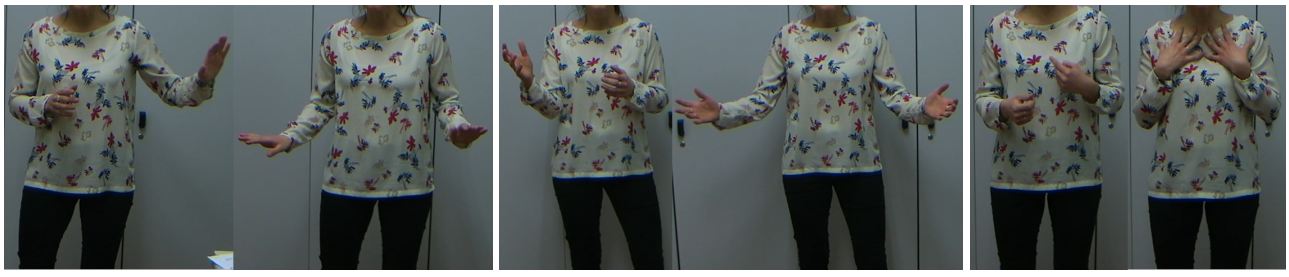
gestures are not conventionalized, have no linguistic properties, and accompany speech. In between the two poles, some other special kind of hand gestures are, for instance, emblems (e.g., the OK gesture in American English) which are partly conventionalized, have some linguistic properties and are used with speech optionally.

In the field of automatic gesture recognition, sign language is the most studied because of the high conventionality of such hands movements (everyone performs the same hand movement in sign language, when signifying a word). Hence, a dataset can be constructed to use machine learning methods effectively. Nevertheless, recognizing more variable gestures still remains a difficult problem. Convolutional neural networks which can achieve very high accuracy in various areas of computer vision, are often used for gesture recognition [Pigou et al., 2017, Cui et al., 2017, Camgoz et al., 2017]. Large RGB-D video datasets have been created for gesture recognition including Italian gestures such as 'andate via' (go away), 'bellissima' (gorgeous), 'd'accordo' (agree), and 'perfetto' (perfect) [Wan et al., 2016]. Although these gestures are more natural than signs they are still emblems, that is, they still have a certain degree of conventionalization. In the datasets, each emblem is performed by many people in almost the same way.

To the best of our knowledge, natural gestures have not been studied extensively. Natural gesture units and phases have been recognized using a support vector machine [Madeo et al., 2016] with scalar velocity and acceleration calculated from hand and wrist positions acquired from a depth sensor used as features. An attempt to estimate spoken words from gestures was made by [Okada and Otsuka, 2017] using an optical motion capture system to acquire motion signals in three-dimensional space, and microphones to capture voice. In general, machine learning methods are often used. However, constructing a dataset for training is an arduous task, and annotating is time consuming and requires linguistic expertise. Gestures change depending on whether the speaker is standing or sitting, or is holding a microphone, and there can be wide natural variations between speakers making the same gesture and between different occurrences of the same gesture made by the same speaker. Hence generating a sufficiently rich and diverse fully annotated training set for recognizing natural gestures via machine learning would be an immense task. Also, whilst optical motion capture systems and depth sensors may be used, when analyzing gestures of a TV news reporter or a speaker on the Internet, only RGB video can be used. Hence, to make this study practical, we aim to recognize natural gestures using only RGB video without machine learning methods.

1.1 Gestures Selected

To test the proposed method, we selected three types of gesture that are not supposed to be conventionalized, nor have linguistic properties: "negation," "palm up," and "me". These were selected because (1) they occur frequently in most kinds of discourse types and contexts (including public speeches), (2) their use is observed in many cultures, and (3) they have a clear semantic and pragmatic meaning. Figure 1 shows examples of the three gestures.



(a) (b) (c)

Figure 1: Examples of the negation (a), palm up (b), and me (c) gestures.

The negation gesture [Kendon, 2004, Calbris, 2003] is performed with the flat palm held downwards or towards the interlocutor, moving laterally. It is derived from actions like sweeping or knocking aside unwanted objects. Such hand movements, often together with head shakes, express the wide semantic theme of negation. The palm up gesture [Müller, 2004] is characterized by palm open upwards, and fingers extended more or less

 ISBN 978-0-9934207-3-3

loosely. It can be performed in isolation or repetitively, following the flow of speech. In general, it serves to present an abstract discourse as a concrete entity. Its most common communicative function is to express an obvious perspective on a topic/entity or to invite the interlocutor to share the speaker's perspective on the topic offered. The me gesture is a special kind of pointing gesture that is directed towards the speaker him/herself. It can consist of one or two hands over the heart, or a simple index finger pointing. It is often used when sharing one's beliefs and ideas, or when talking about something one really cares about.

1.2 Dataset

We recorded three videos in which one of the research team members talked and gestured in a controlled way whilst standing in front of the camera. Each video contains many occurrences of only one type of gesture (either the negation, palm up, or me gesture). While gesturing, the speaker told improvised short stories about cats in Italian. To trigger a coordination of speech and gestures that was as natural as possible, the speaker improvised speech that was associated with the specific kinds of gestures (for example, for the negation gesture, the speaker told stories in the negative (e.g., "the cat did not climb up the tree"); for the me gesture, the speaker told personal stories about her and cats).

The resulting videos were each approximately 4'25" in duration with 25fps. Two rest positions were selected: hands hanging still along the legs (as in Figure 3); hands held still in front of the chest (as in Figures 4 (a) and (b). Each gesture starts from and returns to one of the rest positions. The gestures are performed either with one hand or both hands. The gestures consist of either only one stroke or multiple repeated ones. The speaker performs the gestures with different extension/amplitude and velocity, creating variability between the gestures in terms of motion patterns. Table 1 shows the proportion of frames that are annotated as gestures in each video and the number of gestures in each video.

Table 1: Percentage of frames and number of gestures in each video.

	Ratio[%]	Number
Negation	57.88	36
Palm up	41.92	48
Me	38.25	59

Each video was annotated manually by a linguistic expert using the software ELAN [Brugman et al., 2004]. The labels were *gesture*, *rest* and *other*, where *gesture* denotes anything that happens between two consecutive rest positions; *rest* is any period in which the hands and arms are held in one of the selected rest positions; and *other* denotes anything else (e.g., changing the rest position).

2 Natural Gesture Recognition

We recognize the gestures based on hand trajectory. We assume that the gesturing hand does not change its direction abruptly during any phase of a gesture. The hand position is estimated from the input video using a pose estimation method before recognizing the gestures. Afterwards, the estimated annotation is refined in three steps.

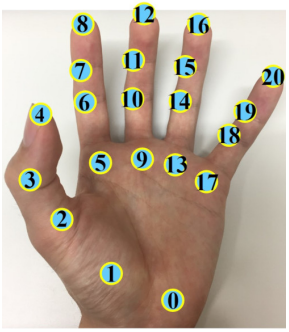


Figure 2: OpenPose can estimate 21 keypoints of a hand.

2.1 Body and Hand Pose Estimation

We detect the body and hand keypoint positions in each frame of the input video using OpenPose[1]. OpenPose is an open source library for real-time multi-person keypoint detection for body [Cao et al., 2017], face, and hands [Simon et al., 2017]. OpenPose is one of the state-of-the-art pose estimation methods. Illustrations of the upper body and hand models generated by OpenPose are shown in Figures 3, 4 (a) and (b).

The 20 keypoints in the hand model other than keypoint 0 (see Figure 2) whose estimated reliability is greater than a specified threshold th_c are averaged in the left (right) hand. Keypoint 0 is ignored as when the

[1]URL: https://github.com/CMU-Perceptual-Computing-Lab/openpose (last accessed: May 9, 2018)

 ISBN 978-0-9934207-3-3

hand is perpendicular to the camera it is often not detected, causing the average hand position to shift from the hand centre to the tip.

As OpenPose occasionally fails to detect all hand keypoints in a frame, any missing averaged hand positions are interpolated using linear interpolation. Interpolated neck, left hip and right hip positions are also acquired in the same way (these are used in future processes). To accommodate large-scale movement of the speaker, such as walking around, for normalization the neck position is subtracted from the hand and hip positions so that the neck position becomes the origin of the coordinate system.

2.2 Hand Trajectory Grouping

The following algorithm is designed to recognize as a gesture those contiguous parts of the hand trajectory that are similar in terms of the direction of motion.

Algorithm Recognizing gestures based on hand trajectory

Input: Hand positions $\boldsymbol{H} = \{\boldsymbol{h}_0, \boldsymbol{h}_1, ..., \boldsymbol{h}_t, ..., \boldsymbol{h}_N\}$, $\boldsymbol{h}_t = \{x_t, y_t\}$
Output: A set of frames annotated as *gesture* $\boldsymbol{g}$

$s = 1, d = 0$
for $t = 1$ to $N - 1$ **do**
  $\boldsymbol{v}_1 = \boldsymbol{h}_t - \boldsymbol{h}_{t-1}$, $\boldsymbol{v}_2 = \boldsymbol{h}_{t+1} - \boldsymbol{h}_t$
  $d \leftarrow d + \|\boldsymbol{v}_1\|_2$
  $\theta = \arccos \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{\|\boldsymbol{v}_1\|_2 \|\boldsymbol{v}_2\|_2} \frac{180}{\pi}$
  if $\theta > th_a$ or $\|\boldsymbol{v}_1\|_2 < th_m$ **then**
    if $d > th_d$ **then** Add from s to t to $\boldsymbol{g}$ **end if**
    $s = t + 1, d = 0$
  end if
end for

The similarity of gesture direction is represented by the angle θ between the two vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ calculated from the hand positions of frames before and after frame t. If θ is smaller than the threshold th_a, frames $t-1$ and t are grouped, and if θ is larger than th_a, the group is divided at frame t.

When the distance from frame $t-1$ to frame t ($\|\boldsymbol{v}_1\|_2$) is smaller than th_m, the group is also separated at frame t. This is because the hand position may move slightly when not gesturing because of the estimation error of OpenPose. As the hand should move to some extent if the group is actually a gesture, the group is annotated as a *gesture* only when the total movement distance of the group d exceeds th_d.

Figure 3: If the hand is inside of one of the white boxes, the hand is judged to be close to the body. l is the hip separation.

2.3 Annotation Refinement

There are three steps in the annotation refinement process.

Connect stroke (CS): If the angle θ between two frames, of which one is annotated as *gesture*, is smaller than th_r ($> th_a$), then the other is also annotated as *gesture*. This process is applied recursively. Firstly, by setting th_a small, only reliable gestures are detected. By using th_r greater than th_a, gestures adjacent to the detected gestures are also detected.

Rest position (RP): Frames that are not annotated as *gesture* while the hand is away from the body are re-annotated as *gesture*. If in a rest position, the hand should be close to the body while it rests. To quantify proximity to the body, if the hand position $\boldsymbol{h} = \{x, y\}$ is within either of the rectangular areas shown in Figure 3, the hand is judged to be close to the body. The two rectangles are defined by the positions of the neck $\boldsymbol{n} = \{x_n, y_n\}$, right and left hip $\boldsymbol{w}_r = \{x_{wr}, y_{wr}\}, \boldsymbol{w}_l = \{x_{wl}, y_{wl}\}$ (please refer Section 2.1), where

$$\{x_{wr} < x < x_{lr} \quad \textbf{and} \quad y_n < y < (y_{wr} + y_{wl})/2\} \quad \textbf{or} \quad \{x_{wr} - l < x < x_{lr} + l \quad \textbf{and} \quad (y_{wr} + y_{wl})/2 < y\}$$

 ISBN 978-0-9934207-3-3

and $l = x_{\mathrm{wl}} - x_{\mathrm{wr}}$. For simplicity, although there is no subscript $_t$ in the formula above, the rectangles are computed at each frame.

Short-term annotation (SA): Finally, as a speaker's behaviour does not in reality change instantaneously, we change short-term annotations that occur between two annotations of the same type. When 1) the length l_{c} of an annotation a_{c} is less than a threshold th_{w} frames 2) the length l_{b} of annotation a_{b} before a_{c} and the length l_{a} of annotation a_{a} after a_{c} are greater than th_{w} frames, and 3) $a_{\mathrm{c}} \neq a_{\mathrm{b}}, a_{\mathrm{b}} = a_{\mathrm{a}}$, then a_{c} is changed to the same annotation as a_{b} and a_{a}.

3 Experiments

Details of the dataset constructed for our experiments are described in Section 1.2. Accuracy (*AC*) (the percentage of frames that were classified correctly) and Jaccard index (*J*) [Wan et al., 2016] were used for evaluation. In Table 1 we see that the proportions of frames that are annotated as *gesture* for ground truth in our three videos range from 38.25% to 57.88%. In completely natural video these proportions are likely to be much lower, resulting in imbalanced class sizes. So we use *J* in addition to *AC* as it does not include true negatives, whereas using *AC* alone could, in general, yield results that are dominated by how well we recognize "a non-gesture". Moreover, we define "trimmed Jaccard index" *tJ* to try to better align our computational measure with the approach of expert linguists when annotating video manually. Considering that the start and end of a gesture are not annotated so precisely when a human annotates the video, *tJ* is calculated by trimming 5% of the frames at the start and end of each gesture. We only used frames that are annotated as *gesture* or *rest*. The thresholds were set as follows throughout all experiments: $th_{\mathrm{c}} = 0.3$, $th_{\mathrm{a}} = 80$, $th_{\mathrm{m}} = 1$, $th_{\mathrm{d}} = 60$, $th_{\mathrm{r}} = 120$, $th_{\mathrm{w}} = 10$.

3.1 Results

Videos were annotated with respect to the movement of left and right hands separately and of both hands together. As described in Section 2, the proposed method recognizes the gestures separately for the right and left hands. So we can evaluate the proposed method for both hands, right hand only, and left hand only. To evaluate both hands, we used logical "or" to integrate the recognition results of the right and left hands, so only when both recognition results were *rest* was the frame classified as *rest*.

Tables 2, 3 and 4 show confusion matrices for each video with both hands. *AC*, *J* and *tJ* for each video with both hands, right hand only and left hand only are shown in Table 5.

Table 2: Negation (both hands).

	$N = 6063$	Predicted *rest*	Predicted *gesture*	
Actual	*rest*	2022	228	2250
	gesture	248	3565	3813
		2270	3793	*AC* 92.15%

Table 3: Palm up (both hands).

	$N = 6162$	Predicted *rest*	Predicted *gesture*	
Actual	*rest*	2890	480	3370
	gesture	28	2764	2792
		2918	3244	*AC* 91.76%

3.2 Recognition of *Hold*

A temporary stop during a gesture is called *hold*. As mentioned in Section 2.3, a stop that is not close to the body cannot be *rest*, so *holds* can be recognized by considering stops that occur outside the rectangular regions shown in Figure 3. The results of adding *hold* annotations to the negation video are shown in Tables 6 and 7. These results are with both hands, so *hold* is identified when one hand is *hold* and the other is either *hold* or *rest*.

 ISBN 978-0-9934207-3-3

Table 4: Me (both hands).

	$N = 6388$	Predicted *rest*	Predicted *gesture*	
Actual	*rest*	3444	401	3845
	gesture	1144	1399	2543
		4588	1800	*AC* 75.81%

Table 5: Accuracy AC[%], Jaccard index J and trimmed Jaccard index tJ for each video with both hands, right and left hand.

Index	Video	Both	Right	Left
AC	Negation	92.15	94.92	93.27
	Palm up	91.76	93.44	93.90
	Me	75.81	80.23	81.40
J	Negation	0.7997	0.7080	0.6996
	Palm up	0.8023	0.7807	0.7942
	Me	0.5228	0.4304	0.4122
tJ	Negation	0.8522	0.7529	0.7456
	Palm up	0.8592	0.8339	0.8519
	Me	0.5736	0.4712	0.4456

Table 6: Negation with *hold* annotation.

	$N = 6061$	Predicted *rest*	Predicted *gesture*	Predicted *hold*	
Actual	*rest*	2049	258	0	2307
	gesture	118	2836	104	3058
	hold	103	499	94	696
		2270	3593	198	*AC* 82.15%

3.3 Effect of Each Process

We examined the effect of including or excluding each of the three processes of annotation refinement described in Section 2.3. A set of experiments were conducted covering all permutations: including only one of the three annotation refinement processes each time; and excluding only one of the three annotation refinement processes each time. The results are shown in Table 8. In Table 8, "+CS", for example, denotes that only the "Connect stroke" refinement process (in Section 2.3) is used; "-CS" denotes that only the "Connect stroke" refinement process is excluded. "Full" denotes that all three processes of annotation refinement are performed, and "None" denotes that none of the three refinement processes is performed. Note that the processes other than annotation refinement are the core of the proposed method. Since the gestures cannot be recognized without them, no experiments excluding the core processes were carried out.

Table 7: Jaccard index J and trimmed Jaccard index tJ of the negation with holding annotation.

	Gesture	Holding
J	0.6689	0.0806
tJ	0.7176	0.0901

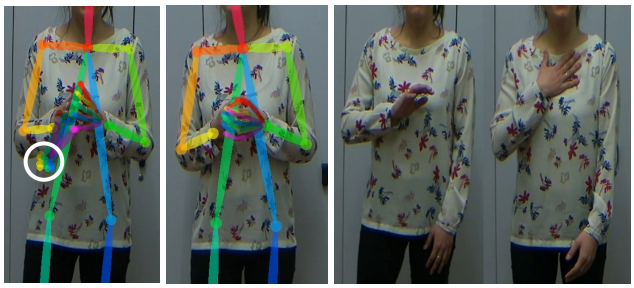
Table 8: Accuracy AC[%], Jaccard index J and trimmed Jaccard index tJ for each video with/without each annotation refinement process. The best and worst scores are highlighted in bold and italic respectively.

		None	+CS	+RP	+SA	-CS	-RP	-SA	Full
Negation	*AC*	*77.19*	85.32	90.73	80.11	91.70	88.60	91.16	**92.15**
	J	*0.5199*	0.6301	0.6842	0.6405	0.7910	0.7526	0.7065	**0.7997**
	tJ	*0.5376*	0.6699	0.7191	0.6688	0.8323	0.7990	0.7533	**0.8522**
Palm up	*AC*	*83.74*	87.15	94.32	87.93	**95.39**	89.76	91.69	91.76
	J	*0.5319*	0.6070	0.6483	0.7378	**0.8481**	0.7795	0.6518	0.8023
	tJ	*0.5871*	0.6487	0.7229	0.7821	**0.9022**	0.8339	0.6984	0.8592
Me	*AC*	74.81	76.10	74.87	*71.52*	71.62	75.81	**76.14**	75.81
	J	0.4063	0.4840	0.4075	*0.3858*	0.3871	**0.5228**	0.4849	**0.5228**
	tJ	0.4488	0.5263	0.4501	*0.4114*	0.4121	**0.5736**	0.5273	**0.5736**
Average	*AC*	*78.58*	82.86	**86.64**	79.85	86.24	84.72	86.33	86.57
	J	*0.4860*	0.5737	0.5800	0.5880	0.6754	0.6850	0.6144	**0.7083**
	tJ	*0.5245*	0.6150	0.6307	0.6208	0.7155	0.7355	0.6597	**0.7617**

 ISBN 978-0-9934207-3-3

4 Discussion

In this section, we analyze the errors. There were two typical errors. The first was caused by the estimation error of OpenPose. This occurred frequently when the speaker was holding hands as shown in Figure 4 (a), where the right hand position was estimated in the white circle. Since it was considered that the hand moved quickly from the actual position to the circled position, this movement was recognized as a *gesture* (in fact, it was *rest*). Although Figure 4 (b) seems to be free from the estimation error of the hand position, when looking at the video, the hand positions are seen to be swapped frequently, fail to be detected, or are jittering. The *rests* tended to be classified as *gesture* in this case. Although these mis-recognitions can be prevented by adjusting the threshold th_{m}, this may result in failure to detect small gestures. Secondly, as shown in Figure 4 (c), cases of *holds* close to the body were easily mis-recognized as *rest*. In this case, the method described in paragraph "Rest position" in Section 2.3 is not effective. As this error occurred frequently in the me gesture, the accuracy of me gesture recognition was lower than that of negation and palm up.



(a) (b) (c)

Figure 4: Typical cases of pose estimation error (a) and (b), and *hold* close to the body (c).

The gesture recognition accuracy for each hand separately is higher than that of both hands together, as shown in Table 5. It is conceivable that this is because the proposed method is designed to perform gesture recognition for each hand. However, J is highest with both hands. This is because the gestures with both hands were found by integrating the classifications of the right and left hands using logical "or", and only the gesture recognition rate is important in J. tJ is about 0.05 higher than J. This suggests that false recognitions are more frequent immediately after the start and just before the end of the gestures. These errors are not critical since the annotation of transitions between gestures and rests can be difficult even for expert linguists.

In Table 6, we see that *rests* were not mistakenly recognized as *hold*. The hands were always near the body for *rests* (stops when the hand was far from the body were annotated as *hold*). Some *holds* tended to be annotated as *gesture*. Theoretically the proposed method does not recognize the *holds* (it just deems a *rest* outside the body area to be *hold*). The proposed method just tries to identify *gestures* and *rests*. Overall, then using *hold* annotations, we see from Table 7 that *gestures* were still reasonably successfully recognized, but *holds* were hardly recognized.

Finally we analyze Table 8. In most cases "None" performed worst, and "Full" performed best or close to best. For palm up, "-CS" performed best. The algorithm recognizes parts of the hand trajectory that are similar in terms of motion direction as *gesture*, and then recognizes other parts that are adjacent to the gesture parts and are somewhat similar in the motion direction to the gesture parts. Although this two-stage approach aims to improve the recognition accuracy, in the case where the gestures can be recognized by only the first stage (the direction of the hand trajectory does not change greatly during the gesture), the second stage would recognize the rests as *gesture*. Comparison of "-SA" and "Full" shows that J and tJ are greatly improved by "SA". This indicates that the "SA" refinement process, which guarantees a certain length of annotations, was effective because J is calculated for each gesture sequence. There was no corresponding improvement in AC, which is calculated for each frame. "RP" also had some effect since the hands were always close to the body. However, "RP" was no meaning for the me gesture, where the hands were always close to the body even during gestures.

5 Conclusion

We proposed a method to recognize natural gestures using only an RGB video without machine learning methods in order to make the proposed method practical. We recorded videos in which a speaker gestured while

 ISBN 978-0-9934207-3-3

talking to test the proposed method. We recognized the gestures automatically based on computation of hand position from a pose estimation method, followed by characterizing the hand trajectory. Recognition accuracies were 92.15%, 91.76% and 75.81% for three natural gestures.

The experiments showed that when the hands moved away from the body during the gestures, the gestures could be recognized with good accuracy, but it was more difficult to recognize other gestures. The proposed method requires use of thresholds rather than using machine learning. For assisting a linguist to carry out video annotation, even if a user needs to adjust the thresholds manually, our goal will be achieved if the user's effort is reduced compared with a fully manual annotation. However, excessive selection of thresholds increases user effort. In the future, we will propose a generalized method with fewer thresholds. We will also expand the range of gesture types and focus on recognizing gesture phases such as preparation and retraction. In this work, the videos used in the experiments were recorded in the same setting and with the same speaker. To verify that the results do not depend on the speaker and the setting, further experiments will be conducted with different speakers and settings, and ultimately on speech videos on the Internet.

Acknowledgments

This work was supported in part by a Grant-in-Aid from the Japan Society for the Promotion of Science Fellows Grant Number 17J05489.

References

[Brugman et al., 2004] Brugman, H., Russel, A., and Nijmegen, X. (2004). Annotating multi-media/multi-modal resources with elan. In *LREC*, pages 2065–2068.

[Calbris, 2003] Calbris, G. (2003). From cutting an object to a clear cut analysis: Gesture as the representation of a preconceptual schema linking concrete actions to abstract notions. *Gesture*, 3(1):19–46.

[Camgoz et al., 2017] Camgoz, N. C., Hadfield, S., Koller, O., and Bowden, R. (2017). Subunets: End-to-end hand shape and continuous sign language recognition. In *ICCV*, pages 3056–3065.

[Cao et al., 2017] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299.

[Cui et al., 2017] Cui, R., Liu, H., and Zhang, C. (2017). Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *CVPR*, pages 7361–7369.

[Dell, 1970] Dell, C. (1970). *Primer for Movement Description Using Effort/Shape*. Princeton Book Company Publishers.

[Kendon, 1980] Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication*, 25(1980):207–227.

[Kendon, 2004] Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

[Madeo et al., 2016] Madeo, R. C. B., Peres, S. M., and de Moraes Lima, C. A. (2016). Gesture phase segmentation using support vector machines. *Expert Systems with Applications*, 56:100–115.

[McNeill, 2008] McNeill, D. (2008). *Gesture and thought*. University of Chicago press.

[Müller, 2004] Müller, C. (2004). Forms and uses of the palm up open hand: A case of a gesture family. *The semantics and pragmatics of everyday gestures*, 9:233–256.

[Okada and Otsuka, 2017] Okada, S. and Otsuka, K. (2017). Recognizing words from gestures: Discovering gesture descriptors associated with spoken utterances. In *FG*, pages 430–437. IEEE.

[Pigou et al., 2017] Pigou, L., Van Herreweghe, M., and Dambre, J. (2017). Gesture and sign language recognition with temporal residual networks. In *CVPR*, pages 3086–3093.

[Simon et al., 2017] Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, pages 1145–1153.

[Wan et al., 2016] Wan, J., Zhao, Y., Zhou, S., Guyon, I., Escalera, S., and Li, S. Z. (2016). Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *CVPRW*, pages 56–64.