Automatic annotation of head velocity and acceleration in Anvil

Jongejan, Bart

Published in:
Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Jongejan, B. (2012). Automatic annotation of head velocity and acceleration in Anvil. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12) (pp. 201-208). European language resources distribution agency.

Automatic annotation of face velocity and acceleration in Anvil

Bart Jongejan
University of Copenhagen, Centre for Language Technology (CST)
Njalsgade 140, 2300 København S, Denmark
E-mail: bartj@hum.ku.dk

Abstract

We describe an automatic face tracker plugin for the ANVIL annotation tool. The face tracker produces data for velocity and for acceleration in two dimensions. We compare the annotations generated by the face tracking algorithm with independently made manual annotations for head movements. The annotations are a useful supplement to manual annotations and may help human annotators to quickly and reliably determine onset of head movements and to suggest which kind of head movement is taking place.

Keywords: Face detection, Automatic discourse annotation, Inter-annotator agreement

1. Introduction

For a human annotator the manual segmentation and annotation of head movements in a multimodal corpus is a time consuming task. Inter-annotator agreement about segmentation is sub-optimal (Navarretta & Paggio, 2010). Automatic discourse annotation of head movements has the potential of making the annotation process swifter and less prone to personal choices. Also, automatic annotations, being based on raw position data without any psychological bias, may result in a physiological description of head movements in terms of velocities and (muscular) forces (acceleration) that is interesting in its own right and that can be compared with high level descriptions in terms of nods, shakes and other descriptors bearing conversational connotations.

2. Background

As reported in (Jongejan, 2010) we have added a face tracker plugin [1] to Anvil (Kipp, 2008), a generic annotation tool for multimodal dialogue. The plugin is based on OpenCV (Bradski & Kaehler, 2008), using the JavaCV [2] programming interface to bridge the gap between OpenCV's C/C++ world and Anvil's Java world. Since our previous report, our algorithms for tracking faces have been much improved.

Earlier, Al Moubayed et al. (2009) have used OpenCV to detect faces. They applied the Lucas-Kanade algorithm to compute velocity as a function of time. By filtering away the low frequency component they obtained a signal that corresponded to e.g. head nods and shakes. Using the optical motion capture system Qualisys, Cerrato & Svanfeldt (2005) obtained automatic annotations for head nods. Their detection algorithm was based on velocity, a minimum number of consecutive frames and, in contrast to the current work, the amplitude of the head movements.

3. Method

In our setup two people are in a dialogue that is filmed by two cameras, one for each participant. The video streams are combined into a split screen video, showing both participants' upper bodies and heads obliquely oriented towards a camera whereas, in reality, they are oriented towards each other. Because of this set-up, we use the OpenCV Haar-based routine for frontal face detection. If, as in our set-up, there are more than one faces in the scene, the user can select the person to analyse by pointing at the person's face and clicking with the mouse. Optionally, the user can instruct the face tracker to stay and wait in the left or right half of the screen during periods when OpenCV cannot detect the face. See fig. 1. In the more common set-up with only one person in the field of view, more or less looking in the direction of the camera, the face tracker automatically finds the person's face and can be run without human supervision and intervention for the full length of a video.

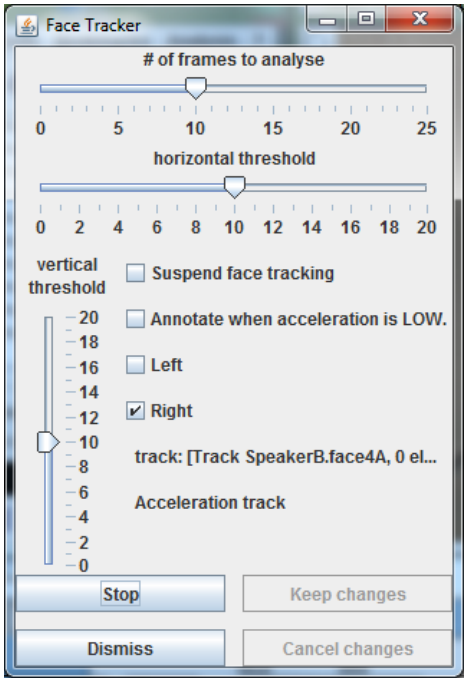

Figure 1. The face tracker window

[1] https://github.com/kuhumcst/Anvil-Facetracker
[2] http://code.google.com/p/javacv

We have analysed the change in time of the face positions. The working hypothesis is that e.g. a head nod not only changes the orientation of a face, but also its lateral and vertical centre position.

We extracted velocity and acceleration vectors that are well correlated with head nods, shakes and other head movements. Each vector is based on a sequence of face positions of up to 25 frames, an analysis window covering a time span in the order of one second. For each frame, the program stores the horizontal and vertical positions of the face in a cyclic buffer containing enough cells to memorize a full analysis window. The mean velocity and acceleration during the window's time span are computed by applying linear regression analysis on the data in the cyclic buffer.

The user can adjust the number of frames in an analysis window (Fig. 1), trading detail in the time domain against statistical reliability. The regression analysis requires at least two or three frames for the computation of velocity and acceleration, respectively. In practice, using much fewer than 8 frames makes the algorithm more sensitive to outliers, which for example can be caused by glitches in the recognition of the face in some frames. Using many more frames than 10 has the effect of desensitising the algorithm for quick, short movements that take place in a fraction of a second. Bursts of acceleration will go unnoticed if the duration of the analysis window is much longer than the bursts.

For every new frame the data for the oldest stored frame are removed from the cyclic buffer and the data for the new frame are added, shifting the span in time of analysed data by just one frame. So although looking at a high number of frames in each an analysis window smears out the data over time, the chosen method does not coarse grain the time domain from the user's perspective: the generated annotations can begin at any frame.

We create annotations for those time spans in which the velocity (or acceleration) is above a set threshold. The user can set thresholds for velocity and acceleration in the horizontal direction as well as in the vertical direction (Fig. 1). An annotation starts at the earliest frame in the earliest analysis window in which the threshold was surpassed, and it ends at the last frame in the last analysis window in which the threshold still was surpassed. If the onset of an annotation would be before the end of the previous annotation, the onset is delayed until the end of the previous annotation, because Anvil does not allow overlapping annotations in the same annotation track. The shortest time span for an annotation is the duration of the analysis window, except when the onset was delayed, in which case an annotation can be as short as two frames.

The video overlay window is used to continuously inform the user about which part of the video is analysed, where the chosen physical quantity (velocity or acceleration) currently is pointing and whether the quantity is below or above the set thresholds. In fig. 2 the person on the right side nods, according to a manual annotation. The current velocity (yellow arrow) points in the "12 o'clock" direction. The red circle (or ellipse) indicates the currently set thresholds for velocity components in the horizontal and vertical direction. Because the arrow reaches out of the red circle, an annotation will be created. The black square delineates the part of the image that is sent to the OpenCV software and is continuously adjusted in size and position. The person on the right in fig. 3 tilts her head, according to the manual annotation. The cyan arrow designates the current acceleration, which is "8 o'clock".

Fig. 4 shows a part of the annotation window. In the top line is a manual annotation: Nod, Repeated. The frame shown in fig. 2 is taken from this event.

Figure 2. Person nodding

Figure 3. Person tilting her head

.

The other annotations are automatically created and each contains three time stamped points. The last two time stamped points contain coordinates indicating the initial and terminal point of the vector that represents the observed quantity when it was at its greatest during the time span of the annotation. The first time stamped point contains the size and direction of the same vector in polar coordinates. For example in fig. 4, (75x12 03:29:48) indicates a vector with size '75' and a clock direction of 12 o'clock, representing any angle in the interval between 345° and 15°. The annotations in the middle are velocities, first upwards (12), then downwards (6). The annotations in the bottom line are accelerations. First up (1), then down (6), up again (11) and finally down (5). The strengths are decreasing until both velocity and later acceleration stay below their thresholds. The yellow marks indicate when the tracked quantity reached its maximum value (the first coordinate of said point) during the time span of the annotation. Notice that all but the first annotation are so much shortened on the left side that their

maxima fall outside the annotations and instead land in earlier annotations.

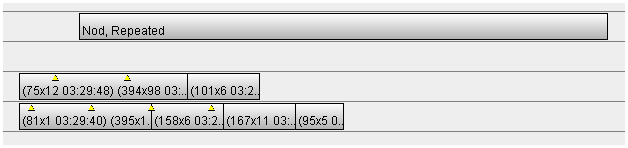

Figure 4. Part of the annotation window

The current hardware and software setup cannot quite attain a real time analysis of every frame in a video. Initially we allowed the analysis software to skip as many frames as needed to keep up with real time video, but after comparing these results with results obtained by frame-by-frame analysis of the same video material we concluded that skipping frames is indeed a bad idea, as already noted by (Matsusaka, 2009).

4. Experiment

As test material for the face tracker we used a video containing about five minutes of dialogue between two persons (Fig. 2). On beforehand, the video was manually annotated by one annotator and checked by a second annotator. In cases of doubt a third person was involved.
We repeatedly run the face tracker for the full length of the dialogue. We did this for each person.
While keeping the duration of the analysis window at 10 frames (about 0.4 seconds), we varied the thresholds between 5 and 14. To give an impression of what these numbers mean, a threshold of 10 for velocity corresponds to a velocity of 0.1 'head size' per second or about 3 cm/s. An acceleration threshold of 10 corresponds to an acceleration of 0.5 'head size' per second squared, or roughly 15 cm/s^2. An acceleration of 15 cm/s^2 sustained in the same direction during 0.4 seconds, starting from rest, results in a velocity of 6 cm/s and displaces the head by just 1.5 cm. If the head already moves at a velocity of 3 cm/s and an acceleration of 15 cm/s^2 counteracts the movement during a time span of 0.4 s, the velocity never gets above the threshold value of 3 cm/s but is instead reversed to the opposite direction. From this we can conclude that, with these settings, which are the default settings, an acceleration annotation without a corresponding velocity annotation is indicative of change of direction of a head movement.

5. Analysis

We have compared the automatically generated annotations with manual annotations. For each video frame in the sequence of almost 8000 frames (5'20"), we checked whether the human annotator and the face tracker agreed or not under the assumption that a specific class of automatically generated annotations was equivalent to a specific class manual of annotations. Because we wanted to compensate for agreement by chance, we chose Cohen's kappa (Cohen 1960) as a measure of agreement.
The classes of manual annotations we looked at were the seven distinct communicative gestures with the head that occurred in our test material: HeadForward, (down-)Nod, Shake, SideTurn, Tilt, Waggle, and HeadOther. The automatic annotations were categorized in 240 distinct classes, each class defined by quantity (velocity or acceleration), clock direction (1-12), and threshold (5-14). For each person we computed 7x240=1680 kappa values. To keep the number of variables manageable, in this analysis we ignored the maximum magnitude of velocity or acceleration during an annotation's time span. We also ignored additional information in the manual annotations, such as whether a head nod was repeated or not.
Also disregarded were the manual annotations for facial expressions and for body posture, although the face tracker isn't insensitive to body postures and even facial expressions. For example, fig. 5 shows a manual annotation for the body posture "BodyDirectionOther, BodyToInterlocutor" that neatly corresponds with automatic annotations for velocity and for acceleration over a range of thresholds: face0V and face0A (the annotation tracks just below the BodyPosture track) have threshold 14, while face9V and face9A lie in the opposite end of the spectrum with thresholds of 5.

Figure 5. A BodyToInterlocutor posture

As there are two persons in our experiments, the left person looking obliquely to the right and the right person looking obliquely to the left, we did the analysis for each person separately. A priori we expected that the mirror symmetry of the set-up might be detectable in the results of the analysis, the roles of e.g. the 5 o'clock and 7 o'clock directions for the left person being swapped to 7 o'clock and 5 o'clock directions for the right person. For example, pure up-down movements such as nods are seen

from an oblique angle by the cameras and will therefore obtain horizontal components that are mirrored with respect to the left person and the right person. Because of the expected differences between the time series for the two persons, we decided not to pool them together.

6. Results

A priori we expected that clock directions of around 6 would correlate with HeadForward and Nod, and that clock directions around 3 or 9 would strongly correlate to Shake and SideTurn. How well does a statistical analysis corroborate our expectations and what has such an analysis to tell us about other clock directions, and are there any statistical clues to decide whether a Tilt, Waggle or HeadOther is occurring?

The following four tables, which are the results in condensed form of the computation of 1680 kappa values., show for each of the two quantities, velocity and acceleration, and for each person in figs. 2-3, which manual head movement annotation is best in agreement with an automatic annotation, given a measured direction of either velocity or acceleration and the threshold as it was set when the measurement was made.

The leftmost column enumerates the clock directions, '1' corresponding to an angle of 30° in a clockwise direction from the vertical (or rather, the interval from 15° to 45°), '2' corresponding to an angle of 60°, and so on. The second column enumerates all thresholds used during the experiment, from low (most sensitive to movement) to high (least sensitive). Ranges of thresholds are put into the same row if all values in the range, according to the statistical analysis, best corresponded to the same manual annotation, 'best' being defined by the highest Cohen's kappa. The third column indicates the lowest and highest kappa measured for the agreement between the manual annotation in the fourth column and the automatic annotation defined by the clock direction in the first column and any of the thresholds in the second.

The predictive power of an automatic annotation seems to vary with the clock directions. As Table 5 shows, some directions correspond to just one head movement whereas others are very sensitive to the threshold and have the full range of thresholds divided in up to seven ranges.

The 6 o'clock direction jumps out as the least ambiguous direction overall. In the case of velocity measurements, this direction indicates HeadForward or Tilt most strongly. In the case of acceleration measurements, this direction indicates a Nod. The 12 o'clock is a contender, especially for the person on the right, who during nodding movements tends to make an upward movement that is strong enough to be noticed at all threshold levels. As to the two other head movements that we had an a priori feeling about, Shake and SideTurn, the picture is less clear. The right person's SideTurns correspond nicely with velocity in the 9 o'clock direction (but not 3 o'clock). For the person on the left this direction is dominated by Shake and HeadForward movements in the velocity domain, while the SideTurn is to be found at 4 o'clock. The left person's Shake movements are better determined by looking for accelerations in the 9 o'clock direction, while for the person on the right there seems not to be a good way to pinpoint the Shake movements, although she makes them.

There are a few negative kappa values in the tables 1-4. They are always very small. If at all, they can only be taken as a weak counter-indication of a head movement.

The tables 1-4 should be interpreted with care. If a direction is shown with only one head movement, it does not mean that no other movements have been noticed to take place in that direction. This just means that the shown head movement statistically is stronger correlated to that direction than all other head movements. The numbers do not disclose how far the shown head movements are ahead of the competitors.

Clock dir.	threshold low/high	Kappa low/high	Head Movement
1	5	0.132	HeadForward
	6	0.091	Nod
	7/9	0.090/0.118	HeadForward
	10	0.068	Nod
	11/14	0.068/0.094	HeadForward
2	5	0.055	Shake
	6/10	0.059/0.072	Waggle
	11/14	0.044/0.064	Tilt
3	5	0.192	Tilt
	6/7	0.100/0.131	Shake
	8/14	0.123/0.166	Tilt
4	5/13	0.038/0.166	SideTurn
	14	-0.007	Waggle
5	5/14	0.103/0.227	HeadForward
6	5/7	0.072/0.096	Tilt
	8/14	0.051/0.074	HeadForward
7	5/11	0.050/0.074	Tilt
	12	-0.005	Waggle
	13/14	0.012/0.015	Nod
8	5/14	0.037/0.118	Tilt
9	5/6	0.109/0.151	Shake
	7	0.102	HeadForward
	8/10	0.100/0.136	Shake
	11/12	0.081/0.092	HeadForward
	13/14	0.052/0.054	Shake
10	5/14	0.072/0.183	HeadOther
11	5/14	0.065/0.161	HeadOther
12	5/9	0.034/0.079	HeadOther
	10/13	0.012/0.016	Nod
	14	0.000	Waggle

Table 1 Velocity. Person on left side.

Clock dir.	threshold low/high	kappa low/high	Head Movement
1	5/6	0.114/0.143	HeadOther
	7/8	0.098/0.116	Nod
	9	0.097	HeadOther
	10/14	0.066/0.111	Nod
2	5/8	0.111/0.133	Tilt
	9/14	0.126/0.211	Waggle
3	5	0.111	HeadOther
	6/7	0.095/0.106	Tilt
	8/14	0.096/0.128	HeadOther
4	5/11	0.047/0.095	SideTurn
	12/14	0.076/0.082	HeadForward
5	5/12	0.067/0.111	Nod
	13/14	0.054	Tilt
6	5/14	0.092/0.126	HeadForward
7	5	0.195	HeadForward
	6	0.043	HeadOther
	7/14	0.051/0.254	HeadForward
8	5	0.084	SideTurn
	6/8	0.068/0.138	Waggle
	9	0.067	HeadForward
	10/11	0.118/0.127	Waggle
	12/14	0.086/0.110	SideTurn
9	5/14	0.104/0.189	SideTurn
10	5	0.077	Waggle
	6/8	0.115/0.125	Tilt
	9/10	0.074/0.084	Waggle
	11	0.069	Tilt
	12	0.074	SideTurn
	13	0.072	Waggle
	14	0.087	SideTurn
11	5/14	0.056/0.176	HeadForward
12	5/14	0.071/0.154	Nod

Table 2 Velocity. Person on right side.

Clock dir.	threshold low/high	kappa low/high	Head Movement
1	5/10	0.052/0.097	Tilt
	11/12	0.023/0.045	HeadForward
	13/14	0.022/0.035	Nod
2	5/14	0.033/0.089	Shake
3	5	0.096	SideTurn
	6/14	0.105/0.162	Tilt
4	5	0.036	Waggle
	6/7	0.061/0.100	Shake
	8	0.088	Waggle
	9/10	0.048/0.053	HeadForward
	11	0.017	Shake
	12/13	0.020	HeadForward
	14	-0.007	Waggle
5	5/14	0.035/0.138	Nod
6	5/14	0.050/0.177	Nod
7	5	0.045	SideTurn
	6/7	0.036/0.049	HeadForward
	8/14	0.014/0.077	HeadOther
8	5/9	0.035/0.113	HeadForward
	10	0.041	Waggle
	11/14	0.033/0.044	Shake
9	5/14	0.101/0.206	Shake
10	5	0.031	HeadForward
	6/9	0.050/0.110	SideTurn
	10	0.074	HeadOther
	11/14	0.074/0.110	HeadForward
11	5/11	0.055/0.120	HeadOther
	12/14	0.010/0.019	Nod
12	5/7	0.112/0.184	Nod
	8/11	0.066/0.104	HeadForward
	12/14	0.043/0.046	Nod

Table 3 Acceleration. Person on left side.

7. Discussion

We primarily used Cohen's kappa as a measure to rank mappings. For that purpose, their absolute values were of no relevance at all. On the other hand, Cohen's kappa expresses inter-coder agreement in an absolute sense, and therefore we must understand why, overall, the kappa values were very low. By knowing the main reasons why they are low and seeing viable ways to get control over these factors, we improve our confidence in Cohen's kappa as a reasonable measure of agreement between manual and automatic annotations:

(1) We have to do with annotators with very different capabilities. The machine seems to be good at pinpointing the onset of a head movement, whereas the human is better at observing the relatively long aftermath of a head movement, when the acceleration and velocity, after an initial burst, already have dropped below the machine's threshold. Temporarily lowering the threshold when a head movement already has been detected, may prolong automatic annotations and improve Cohen's kappa.

lock dir.	threshold low/high	kappa low/high	Head Movement
1	5/14	0.094/0.128	HeadForward
2	5/6	0.116/0.123	Tilt
	7	0.090	Waggle
	8	0.084	Tilt
	9	0.059	Waggle
	10	0.059	Tilt
	11/12	0.071/0.083	Shake
	13/14	0.097/0.101	Waggle
3	5/6	0.062/0.082	HeadOther
	7	0.070	Shake
	8/9	0.082/0.089	HeadOther
	10/11	0.104/0.106	Shake
	12	0.067	HeadOther
	13/14	0.065/0.074	Shake
4	5/10	0.057/0.121	SideTurn
	11/13	0.092/0.098	Waggle
	14	0.119	Shake
5	5	0.024	HeadForward
	6/7	0.024/0.065	Nod
	8/9	0.072/0.081	HeadForward
	10	0.063	Nod
	11/12	0.068/0.074	HeadForward
	13/14	0.045	Nod
6	5/14	0.102/0.253	Nod
7	5/8	0.041/0.092	Waggle
	9/14	0.064/0.114	HeadForward
8	5/13	0.057/0.118	SideTurn
	14	0.038	Shake
9	5/7	0.126/0.148	HeadOther
	8/9	0.154/0.181	Shake
	10	0.129	HeadOther
	11	0.142	Shake
	12/14	0.143/0.182	HeadOther
10	5	0.094	SideTurn
	6/7	0.060/0.083	HeadForward
	8	0.084	Waggle
	9/10	0.075/0.100	SideTurn
	11	0.098	HeadForward
	12/13	0.131	SideTurn
	14	0.123	Waggle
11	5	-0.002	HeadOther
	6	0.014	Nod
	7	0.033	HeadOther
	8/13	0.031/0.057	Tilt
	14	0.042	SideTurn
12	5/14	0.116/0.226	Nod

Table 4 Acceleration. Person on right side.

(2) Even in theory, the classes of manual annotations and those of automatic annotations presented in this paper cannot be mapped onto each other. Most head movements are phrases consisting of several phases. There may be a good correlation between individual phases and automatic annotations. There may also be a good correlation between certain phase transitions and certain automatic annotations, but it is not expected that any movement consisting of three or more phases closely correlates to any automatic annotation. Complex annotations, composed of two or more automatic annotations, may show a much better agreement with manual annotations.

(3) The human annotator only annotated those head movements that were considered to be communicative gestures. The face tracker does not make a distinction between communicative gestures and non-communicative head movements. It may also be the case that human annotators not always want to include a preparation phase in a communicative gesture, such as the slight upward movement before a down-nod. Whereas it may be difficult or impossible to learn the face tracker to skip the non-communicative movements altogether, it seems possible to learn the face tracker to delete certain automatic annotations when found in specific constellations of automatic annotations.

(4) In our analysis, each video frame was considered an individual case, irrespective of preceding and following frames. The fact that the head movement data in an uninterrupted temporal sequence of frames can conglomerate into a single annotation that can overlap with a manual annotation, is not taken into account. Because of this, a disagreement attributable to a true recognition error (e.g. noise in the video signal) and a disagreement attributable to different ‘opinions’ about the precise onset of a head movement are weighted equally in the computation of Cohen’s kappa, resulting in a pessimistic estimation of the agreement between manual and automatic annotations.

(5) Even the human annotators did not agree excellently with a Cohen’s kappa around 0.71 for head movement segmentation and annotation (Navarretta & Paggio, 2010), lowering the bar for the face tracker.

The results presented here are as closely based on raw data as possible. To see the effects of some algorithmic massaging, we also performed the analysis after coarse graining the time domain in chunks of 10 or even 25 frames (0.4 s – 1 s). However, although we obtained better Cohen's kappa values, we lacked a theoretical sound motivation for coarse graining.

	Table 1	Table 2	Table 3	Table 4
Clock	Veloc. L	Veloc. R	Accel. L	Accel. R
1	5	4	3	1
2	3	2	1	7
3	3	3	2	6
4	2	2	7	3
5	1	2	1	6
6	2	1	1	1
7	3	3	3	2
8	1	5	3	2
9	5	1	1	5
10	1	7	4	7
11	1	1	2	5
12	3	1	3	1

Table 5. Number of ranges in each direction

We did not consider percent agreement as a meaningful measure, because even if the face tracker defected and had not created any annotations at all, percent agreement would still be fairly high, as e.g. the number of frames where a person was nodding would be low in comparison to the number of frames where the same person was not nodding, thus agreeing for most of the time with the silent face tracker.

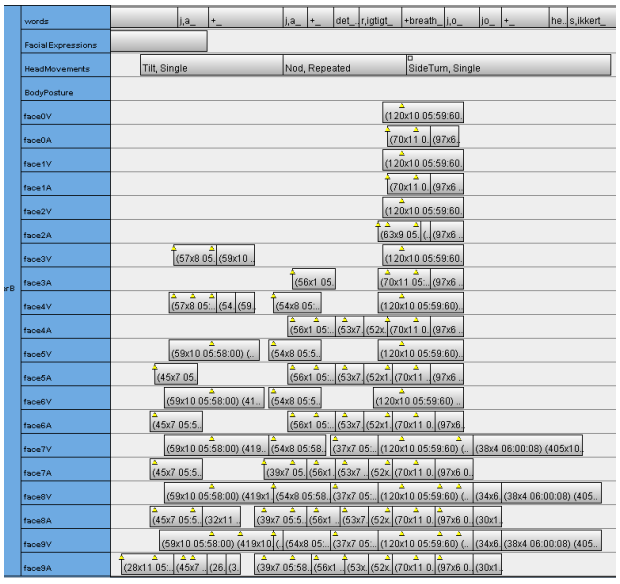

Figure 6. A succession of head movements

Our numerical analysis disregards a number of phenomena that are readily visible when scrolling through the Anvil's annotation window. Here we mention a few.

There are many successions of manual annotations without intervening pauses in the material. Fig. 6 shows a succession of annotations for head movements (a Tilt, a Nod and a SideTurn) together with automatic annotations for velocity and acceleration at ten thresholds levels. Visual inspection immediately tells us that each head movement is automatically detected. At the highest threshold level, only the SideTurn is detected. As the threshold is lowered the Tilt and the Nod are detected, the former by the velocity transgressing the threshold and the latter by the acceleration transgressing the threshold.

The maxima in the annotations in the acceleration tracks always precede the maxima in the velocity annotations. Furthermore, the Tilt and the SideTurn have velocity annotations that continue beyond the end of the acceleration annotations (if they are there at all). During the Nod (which is repeated) it is the other way around: the accelerations go on for a longer time than the velocities: the acceleration directions switch so quickly that the velocity has not time enough to build up. In none of the tracks the automatic annotations continue until the end of the manual annotation. This is the rule rather than an exception.

Visually, there seems to be a good correlation between onset and, sometimes, end of manual annotations and automatically created annotations. Moreover, the automatic acceleration annotations often seem to start a few frames earlier than the manual annotations, as though the human annotator is mostly looking at movements and not at the forces that cause the movements. This discrepancy might be an artefact of the way the program chooses to define the beginning and end of an annotation. Since as mentioned earlier, an annotation covers at least as many frames as are needed to fill an analysis window, by increasing or decreasing this number we can to some degree influence the width of the generated annotations, even without adjusting the sensitivity thresholds. Therefore the discrepancy between onsets of manual and automatic annotations might disappear if we set the duration of the analysis window low enough. This explanation can easily be refuted, however, because, as illustrated by the yellow marks in fig. 4, the maximum value of the chosen quantity during the time span of the whole annotation often lies very close to the beginning of the annotation, well to the left of the start of the manual annotation. Whereas the exact time of the start of an annotation is dependent on program settings like duration of analysis window and thresholds (see figs. 5 and 6), the time of the maximum value of velocity or acceleration is a function of the measured positions of the head during that window. There is no obvious way to explain that time as an artefact.

The last regularity that visual inspection learns us and that goes undetected in the statistical analysis is that a velocity annotation normally is accompanied by two acceleration annotations: one to initiate the movement and the second to stop it. The exceptions to this rule are also interesting. Some movements start and stop so slowly that they are under the thresholds set for acceleration annotations. Such velocity annotations can for example correspond to

movements of the whole body, which normally involve moderate accelerations of the head. The opposite is also possible: a complicated movement along a non-linear path not only involves forces that initiate and stop the movement, but also accelerations that can have strong components perpendicular to the direction of the velocity and that change the direction of the velocity rather than its magnitude. Another situation where velocities can go undetected while accelerations are detected is when acceleration reverses direction so quickly that the velocity has not had time enough to build up to a level that transcends the thresholds for velocity. Such annotations are indicative of short nods and shakes.

8. Conclusion and future work

Automatically created annotations for head velocity and acceleration correlate well with manual annotations of head nods, shakes and other head movements. The onset of the automatically created annotations tends to be a few frames earlier than their manual counterparts. As a rule, manual annotations continue for a considerable longer time than corresponding automatic annotations.

Because of the technical difficulty of keeping up with real time video, suggestions have been made that the automatic annotation of video for face movements, including both face recognition and motion analysis, be performed "off-line". Software all written in C++, directly interacting with the OpenCV algorithms, would certainly be much faster. However, we have found out that observing the play of arrows on screen as the analysis tugs its way through the video – arrows that indicate the current velocity or acceleration – gives new insights that we quite likely would have missed if the analysis had taken place in a batch job without somebody looking. For example, whereas currently a velocity or acceleration annotation stores the direction and size of the largest velocity or acceleration vector occurring during the time span of the annotation, these vectors are seen in many more directions as the analysis takes place. In the case of a straight head shake, the velocity may build up, reach a maximum and decrease, all taking place in the same general direction. But there are many movements where the velocity vector (and the acceleration vector, for that matter) makes a sweeping movement, changing direction over a very wide angle. This observation inspires to implement algorithms that do more right to these movements – typically nods – than the rectilinear approximations offered by velocities and accelerations in a Cartesian reference system. As this phenomenon is currently not taken notice of by the software, it would perhaps have gone undetected if the analysis had taken place in batch mode.

The statistical analysis of a single 5-minute video of a conversation between two people has learned that there are no threshold values that are optimal for detecting all kinds of head movements. The automatic categorization of detected head movements as Shakes, Nods, Tilts and so on can be done, but only with a fair amount of uncertainty. Using machine learning, improvements will be sought by taking into account that many head movements correspond to two or more adjacent automatic annotations in both the velocity and acceleration domain. Reliably establishing these mappings between such complex automatic annotations and their manual counterparts will require the analysis of many more manually annotated dialogues.

9. Acknowledgments

This research has been supported by the Danish Council for Independent Research in the Humanities and by the NOMCO project (http://sskkii.gu.se/nomco/), a collaborative Nordic project with participating research groups at the universities of Gothenburg, Copenhagen and Helsinki. The project is funded by the NOS-HS NORDCORP programme under the Danish Agency for Science, Technology and Innovation.

10. References

Al Moubayed, S.; Chetouani, M.; Baklouti, M.; Dutoit, T.; Mahdhaoui, A.; Martin, J-C.; Ondas, S.; Pelachaud, C.; Urbain, J. and Yilmaz, M. (2009). Generating Robot/Agent Backchannels During a Storytelling Experiment. In *Proceedings of (ICRA'09) IEEE International Conference on Robotics and Automation*. Kobe, Japan.

Bradski, G., Kaehler, A. (2008). *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly

Cerrato, L., Svanfeldt, G. (2005). A method for the detection of communicative head nods in expressive speech. In Allwood, J., Dorriots, B. & Nicholson, S. (Eds.), *Gothenburg papers in Theoretical Linguistics 92: Proc. from The Second Nordic Conference on Multi-modal Communication*, Gothenburg University, Sweden, pp. 153-165.

Cohen J (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*, pp. 37–46.

Jongejan, B. (2010). Automatic face tracking in Anvil LREC - Workshop *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (LREC 2010)*

Kipp, M. (2008). Spatiotemporal Coding in ANVIL. In *Proceedings of the 6th international conference on Language Resources and Evaluation (LREC-08)*

Matsusaka, Y.; Katagiri, Y.; Ishizaki, M. and Enomoto M. (2009). Unsupervised Clustering in Multiparty Meeting Analysis. In M. Kipp, J.-C., Martin, P. Paggio & D. Heylen (Eds.), *Multimodal Corpora. From Models of Natural Intercation to Systems and Applications*, Berlin Heidelberg: Springer Verlag, pp. 93-108

Navarretta, C., Paggio, P. (2010). Classification of Feedback Expressions in Multimodal Data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden, Juli 1116, 2010, pp. 318-324