HAL
open science

Automatic annotation of facial actions from a video record: The case of eyebrows raising and frowning

Stéphane Rauzy, Aurélie Goujon

To cite this version:

Stéphane Rauzy, Aurélie Goujon. Automatic annotation of facial actions from a video record: The case of eyebrows raising and frowning. Workshop on "Affects, Compagnons Artificiels et Interactions", WACAI 2018, Magalie Ochs, Jun 2018, Porquerolles, France. 7 p. hal-01769684

HAL Id: hal-01769684

https://hal.science/hal-01769684

Submitted on 18 Apr 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Automatic annotation of facial actions from a video record: The case of eyebrows raising and frowning

Stéphane Rauzy
Aix Marseille University, CNRS,
Laboratoire Parole et Langage UMR 7309
Aix-en-Provence,France
stephane.rauzy@lpl-aix.fr

Aurélie Goujon
Aix Marseille University, CNRS,
Laboratoire Parole et Langage UMR 7309
Aix-en-Provence,France
aurelie.goujon@lpl-aix.fr

ABSTRACT

Recent advances in the field of computer vision and machine learning have given birth to a generation of softwares enable to detect and track a face along a video record and eventually to measure its internal facial movements. We investigate herein on a specific task, the detection of eyebrows raising and frowning facial actions, the potential of these softwares on spontaneous, in-the-wild video corpus. We propose a tool which allow to extract eyebrows raising and frowning from the output of two state-of-the-art facial behavior analysis softwares (OpenFace and IntraFace). The evaluation performed on our manually annotated in-the-wild video corpus suggests that the tool can be used with benefits for automatic annotation purpose.

Keywords

Facial expression, machine learning, eyebrows movement, corpus, gold standard, automatic annotation

1. INTRODUCTION

The multimodal aspect of human-human interaction is of fundamental importance if one want to model conversational interaction involving humans or virtual agents. During conversation, gestures (hand gestures, head gestures and facial gestures) are generally produced spontaneously along with speech. Many authors have studied the link between hand gestures and discourse structure (see for example [18]). Recurrent features appear which reveals a discourse organization. Some gestures such as hand gestures, head movements and facial expressions can be considered as co-verbal gestures: they occur during speech and cannot be analysed without speech. The distinction between emotional facial gestures and conversational facial gestures rest on four differences: conversational facial gestures are context dependent, speech dependent, without stereotype, and they appear in a social process [5].

Among facial expressions, eyebrow movements are closely connected to speech. At the prosodic level for example, [6] found a strong correlation between eyebrow movements and accentuating intonation contours as well as fundamental frequency. At the conversational level, [1] showed the multimodal aspect of feedback responses (verbal discourse markers 'mh', 'yeah', 'oh' and gestural head nod, eyebrow movement, smile...). [7] made clear that eyebrows raising action structures the start, the continuity and the end of a topic in a conversation. More precisely, eyebrow movements have been predicted to occur more frequently at the start of a new segment in the structure of the dialogue [12]. In a previous study, we also noted that eyebrow movements are associated with the beginning of a new speaking turn in French [15].

The study of the multimodal aspect of conversation requires in practice large corpus containing for each modality the annotations specific to the domain studied.The manual annotation of gestures (hand gestures, facial expressions, ...) is found to be a time-consuming task and put some practical limits on the size of manually annotated corpus available. On the other hand, a large amount of data is needed in order to inventory the various phenomena produced at the interface between the domains under consideration. Any alternative solution to the manual annotation task is therefore welcome.

Automatic analysis of facial expressions from a video record implies several steps of treatment (see for example [17] for a recent survey): a pre-processing of video images in order to detect and track the face and its characteristic facial landmarks all along the video capture, the extraction of features describing for example atomic facial muscle actions (i.e. the Action Units (AUs) of the Facial Action Coding System (FACS) [10, 11]) and a final step allowing to automatically detect facial actions based on the measured features.

Various solutions based on different techniques and algorithms have already led to a bunch of distributed softwares (see table 1 of [4] for an overview and a comparison of the respective characteristics of the available tools). The today challenge concerns how facial behavior analysis softwares do perform on in-the-wild videos recording spontaneous facial expressions (see [8, 17]).

The term 'in-the-wild' is used by the computer vision communauty to describe any realistic settings where the captured face may be far from the frontal head pose, may undergo abrupt head motions, may be masked due to partial occlusions and may be subject to varying illumination conditions. The term 'spontaneous facial expressions' stands for the natural facial expressions anybody experiments during everyday-life social interaction, in contrast with facial expressions resulting from posed emotion played by an actor for example.

The aim of the present paper is to investigate whether the current technology applied to such spontaneous and in-the-wild video records is able to produce good enough automatic annotations of facial actions. To this end the concrete case of eyebrows raising and frowning will be considered. The paper is organised as follows. We present in a first section

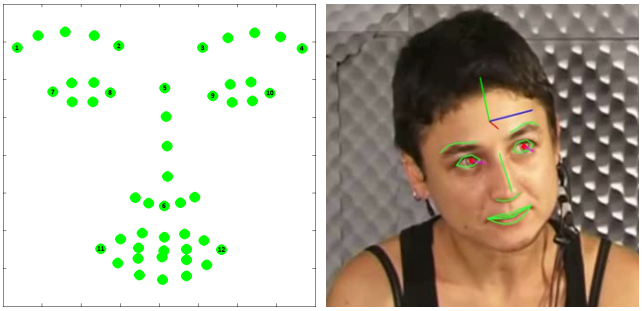

Figure 1: Left panel: The position of the facial landmarks used by the IntraFace software (picture borrowed from the IntraFace documentation manual). Right panel: A frame capture of a processed video showing the landmarks position (green lines), head pose (the projected blue-green-red trihedra) and gaze estimation (red points and lines).

the IntraFace and OpenFace outputs, two state-of-the-art facial behavior analysis softwares, which will be pipelined as the input of our analysis. In a second section, we define the eyebrows raising and frowning actions based on these measurements and describe the method allowing to automatically extract them from the data. An evaluation of the method is then performed on a gold standard corpus. The last section is devoted to discussion and concluding remarks.

2. FACIAL BEHAVIOR ANALYSIS SOFTWARES

2.1 The IntraFace software

The IntraFace software [9] makes use of state-of-the-art algorithms for features detection and tracking (i.e. Supervised Descent Method [22]), head pose estimation, facial attribute recognition, multi-person face detection and analysis, and facial expression recognition. In its publicly available version, the output of IntraFace furnishes the variation in time of the coordinates of a subset of facial landmarks at specific location (eyes corner, nose tip, ... see figure 1), head pose angles and other quantities of interest. The internal facial movement of the proposed landmarks can be retrieved from the output by reconstructing the head model (i.e. the 3-dimensional positions of the landmarks in the head rest frame in absence of internal facial movement) as well as the scale factor function of time which measures the variation of the distance between the head and the camera (reconstruction formulae and other technical issues will be found in Rauzy & Goujon, in preparation).

2.2 The OpenFace software

The OpenFace toolkit [4, 2] is an open source project which proposes the full capabilities of facial behavior analysis tool: head tracking, facial landmark detection, head pose estimation, facial action unit recognition and eye-gaze estimation. OpenFace implements a Constrained Local Model algorithm (CLM-Z, [3]) for facial features tracking. Illustration of the OpenFace processed video is shown figure 2. Reconstruction of the facial landmark movements corrected from head rotation and global head translation is obtained by fitting a head model to the OpenFace output data.

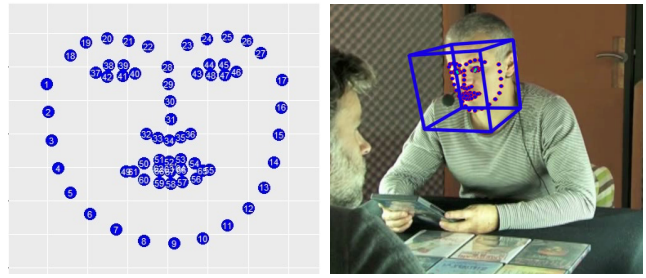

Figure 2: Left panel: The position of the facial landmarks used by the OpenFace software. Right panel: A frame capture of a processed video showing the landmarks position and head pose (the projected blue cube edges).

Figure 3: Examples of configuration with missing measurements. Top left and right panels: The face pose is too far from the frontal pose. Bottom left and right: Partial face occlusion.

2.3 Working on in-the-wild videos recording spontaneous facial expressions

Although OpenFace and IntraFace are state-of-the-art softwares for analysing facial behavior, it remains to evaluate how do they cope in practice in realistic settings with spontaneous facial expressions. Some examples of these configurations are presented figure 3. The top panels show cases where the head pose is too far from the frontal pose, implying the track loss of the head for both softwares, and thus missing measurements at the landmark positions level. Bottom panels of figure 3 show cases of partial occlusions (occlusion due to the second participant in the left panel, self-occlusion for the right panel). Here again, the head track is momentaneously lost by the softwares.

A second and maybe more problematic type of errors arises when the software affects a wrong position to the head. This is illustrated figure 4 for the IntraFace output but the OpenFace software suffers as well of this kind of problems. Some heuristics have to be developped in order to detect and discard these kind of spurious measurements.

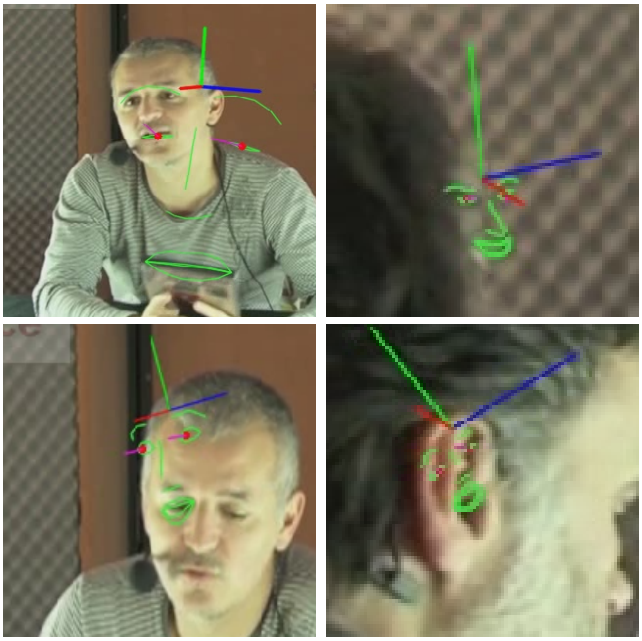

Figure 4: Examples of IntraFace output with problematic measurements.

3. AUTOMATIC DETECTION OF EYEBROWS RAISING AND FROWNING

3.1 Our corpus enriched with eyebrows raising and frowning manual annotations

Eyebrow raises and frowns were manually annotated on some extracts of the Aix MapTask and the Aix-DVD corpus [13, 14, 19]. The characteristics of the 6 corpus extracts are summarized table 1. Our extracts selection has a total duration of 2h 50mn and contains 431 eyebrows raising and 142 eyebrows frowning. The scene configurations (camera positions, illumination conditions, ...) for each extract are presented figure 5.

Manual annotation of eyebrows movements were performed by one expert used to the Facial Action Coding System (FACS, [10, 11]). The task has required to watch the video at a slowed rate to find the exact frames starting and ending the eyebrow movement time interval. Eyebrow raising action is characterized by a gradual elevation of the eyebrows along the vertical axis followed by a fall allowing the eyebrows to go back to their neutral position. Frowning action implies movements on both the horizontal and the vertical axis. Eyebrows move toward each other along the horizontal axis and a line appears between them. Eyebrows action usually involves both eyebrows but single eyebrow movements are not rare as well.

3.2 Definition of the eyebrows frowning and raising functions

Eyebrows frowning and raising actions are described in the Facial Action Coding System (FACS, [10, 11]) as a combination of component movements associated to their underlying muscles (e.g. inner brow raiser (AU1), outer brow raiser (AU2), brow lower (AU4)). Herein frowning and raising actions will be detected by the displacement they induce

Figure 5: The scene configuration for the 6 corpus extracts enriched with eyebrows raising and frowning manual annotations.

for the specific facial landmarks associated with eyebrows. Figure 6 shows the position of the 12 facial landmarks reported by the IntraFace software output. The displacement of landmarks P01, P02, P03 and P04 associated to left and right eyebrow corners of each eye will be of particular interest. They are given by the landmark position residuals in the head rest frame, i.e. the movement corrected from the head rotation and of the head translation bulk motion.

The average displacement for each landmark is herein obtained by comparing in the gold standard corpus the position changes in the 431 eyebrows raising areas (and respectively in the 142 frowning areas) relative to the areas with no eyebrows movement. Figures 7 and 8 show the distribution of these residuals projected respectively on the horizontal

Table 1: Characteristics for each video: duration of the record in second, percentage of frames without IntraFace measurements (%NA), percentage of frames with problematic measurements (%S) and number of manually annotated eyebrows raising and frowning.

Corpus	Duration (s)	%NA	%S	#Raise	#Frown
DVD_AG	2001.04	14.7	7.8	196	18
DVD_YM	1988.44	16.8	1.9	128	82
MTX_AG	2033.88	3.6	4.0	70	6
MTX_YM	2034.88	3.0	2.9	27	33
MTX_AP	1074.68	1.9	4.9	3	0
MTX_JA	1075.84	9.2	2.9	7	3
ALL	10 208.76	8.7	4.0	431	142

Figure 6: Facial positions of the 12 landmarks available from the IntraFace software output.

X and vertical Y axis for the 3 relevant conditions: areas with eyebrows raised, areas with eyebrows frowned and areas with no eyebrows activity. It is for example clear from figure 7 that the horizontal displacement of residual is significantly positive for landmarks positionned at eyebrows corners (from P01 to P04) when raising action occurs (not surprisingly, the landmarks associated with eyebrows corners 'raise'). Herein, the composite raising and frowning functions will therefore be defined as a linear combination of the normalized residuals according to the results obtained on our gold standard manually annotated.

3.3 Automatic detection of eyebrows frowning and raising

Once the raising and frowning functions have been defined, we are left with the problem of detecting temporal variation associated with their corresponding facial activity. A convenient way to solve the problem is to work in the wavelet space. Wavelet transform is a mathematical tool allowing a simultaneous analysis of the positional and scaling properties of a function (see for example [16]). Figure 9 illustrates how the search of a raising or frowning area (i.e. the time interval delimiting the beginning and the end of the action) can be replaced by the search of local maxima in the two dimensional wavelet space.

The method we propose consists therefore in analysing the local maxima of the raising and frowning functions in the wavelet space. We apply herein a standard technique issued from signal processing. The local maxima can be sorted with respect to their signal-to-noise ratio, the higher the value of this signal-to-noise, the higher is the probability to correspond to a true action. This remark allows us to associate a score to each maxima (from 'A' to 'E' for example), from the 'A' value where the signal-to-noise ratio is high and the probability to detect a real raising and frowning action is strong, to 'E' class where the probability to detect an action becomes small. Class 'N' will finally be associated with time intervals where probability of eyebrows activity is low (say that for a signal-to-noise ratio smaller than unity). The time intervals where the raising or frowning functions can not be evaluated because of missing or discarded measurements (see

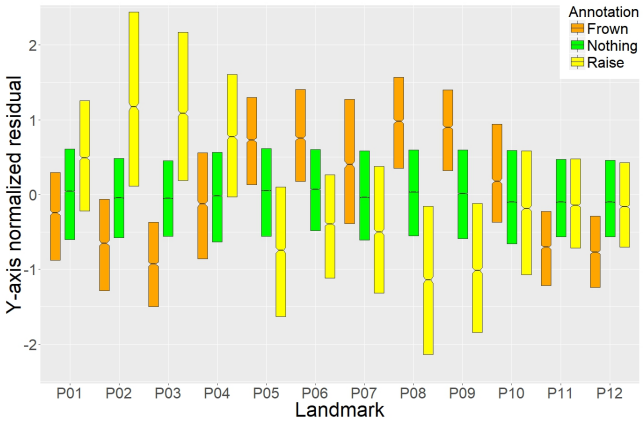

Figure 7: For each landmark, boxplots of normalized residuals along the vertical Y-axis of the head rest frame for the three annotation conditions: eyebrows frowning (in orange), eyebrows raising (in yellow) and without eyebrows movement (in green). The inside band mark shows the median of the residuals distribution. Lower and upper hinges of the boxplot correspond respectively to the first and third quartile of the distribution (i.e. the 25th and 75th percentiles).

Figure 8: Same as figure 7 for the normalized residuals along the horizontal X-axis of the head rest frame.

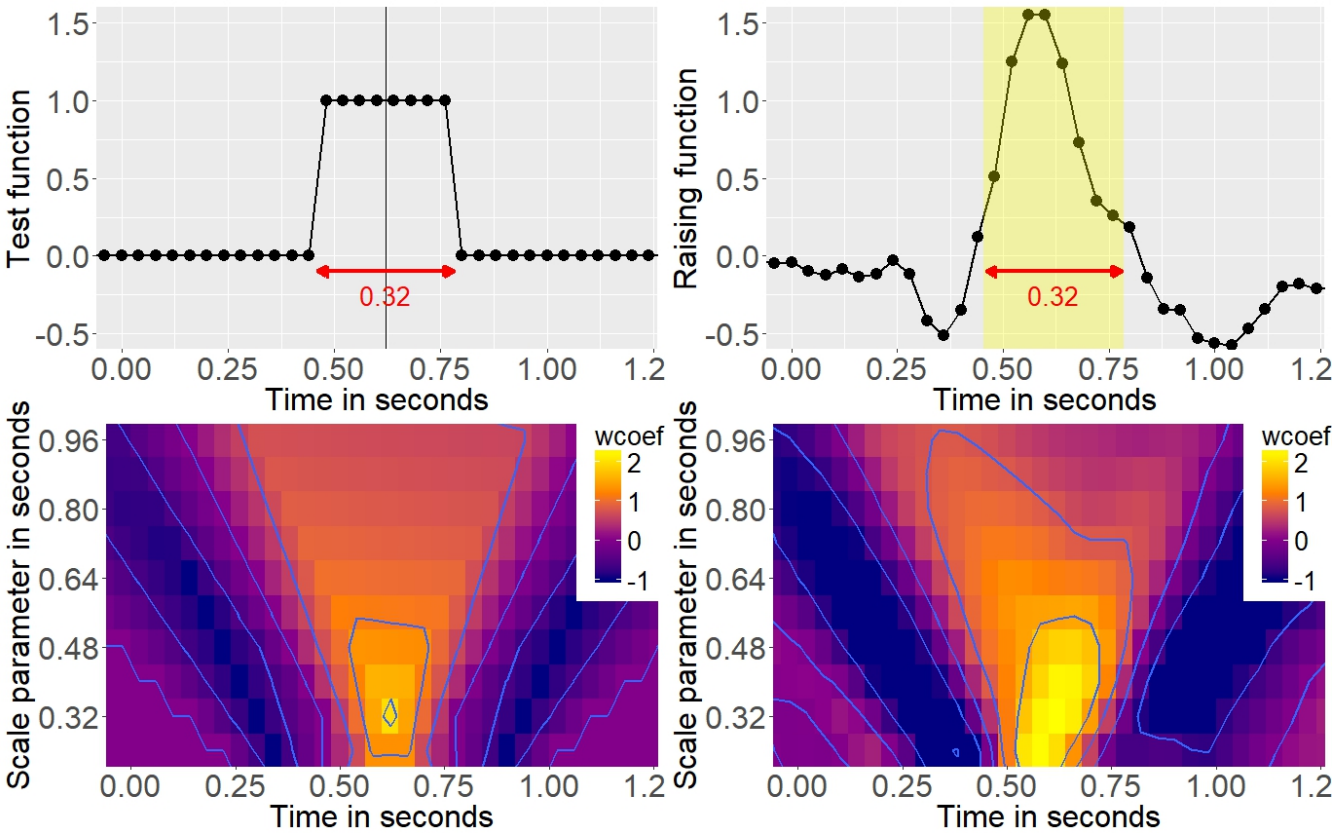

Figure 9: Illustration of the method used to extract automatically time intervals with eyebrows raising or frowning. Top and bottom left panels: A test function and its wavelet transform in the two dimensional space defined by the time and the scale parameter. The maximal value of the wavelet coefficients (yellow color) is found at the time corresponding to the center of the test function interval and with a scale parameter equals to the interval length. Top and bottom right panels: The same figures for a raising function extracted from our data. The yellow area in the top panel delimits the manually annotated eyebrows raising action interval. The search of local maxima in the wavelet space allows to find the center and the duration of eyebrows raising and frowning time intervals.

section 2.3) are annotated with class 'X'. The output of our annotation tool finally consists in a serie of adjacent time intervals covering the whole corpus and annotated with 'X' or 'N' labels or from 'A' to 'E' for intervals potentially associated with an eyebrows action. In that case, the type of action (i.e. 'Raise' or 'Frown') is also mentionned. An output (Elan [21] file format) of our tool is illustrated figure 10.

3.4 Evaluation

Table 2 presents the evaluation of the method on our gold standard corpus manually annotated. The evaluation makes use of the notion of precision and recall initially introduced in the field of information retrieval. The evaluation metric is defined by comparing the automatic output consisting of predicted adjacent time intervals labeled with eyebrows raise or frown action for class 'A' to 'E' intervals, and with no label for 'N' or 'X' classes, versus the manually annotated intervals of the gold standard. The metric is based on the predicted ouput intervals belonging to classes 'A' to 'E'. A predicted interval is considered to match an observed raising or frowning gold standard one if the two intervals overlap (and notwithstanding their overlap ratio nor the action type). The evaluation is conducted class by class.

The evaluation table for the automatic annotations produced by our tool pipelined with the IntraFace software output is presented table 2. On the total of 573 annotated eyebrow movements, 25 stand in 'X' areas with missing measurements. The 548 remaining annotations are distributed in classes 'A' to 'E' predictions except 14 which are classified in the 'N' class. For each class, the precision runs from 0.59 for class 'A' (59% of the class 'A' intervals match with an observed eyebrows movements) to 0.02 for class 'E'. The column named 'match' indicates the rate of success for the predicted eyebrows action type ('Raise' or 'Frown'), e.g. a ratio of 1 indicates that there is no mismatch between the observed and the predicted action type. The total time spanned by the intervals can be found in the 'duration' column. For example the 171 class 'A' predictions span a total duration of 121.98 seconds corresponding to about 1.2% of the whole corpus duration. The recall is the fraction of the observed actions successfully predicted for a given class, e.g. class 'A' contains 18% of the observed eyebrows actions.

The table also proposes the cumulative counts, the preci-

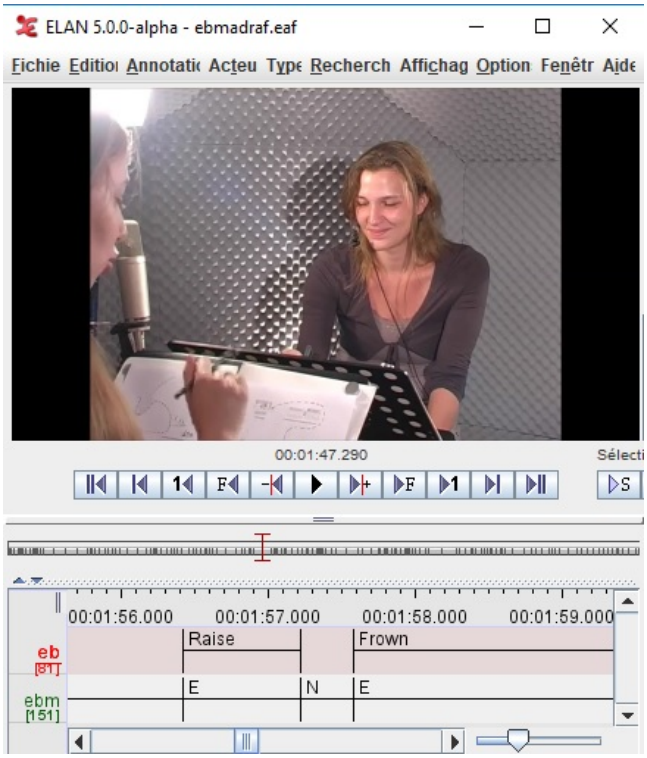

Figure 10: Illustration of the output of our annotation tool (Elan software format).

sion and recall measures over the set of merged classes 'AB', 'ABC', 'ABCD' and 'ABCDE'. These quantities are usefull in order to evaluate the coverage properties of the automatic tool. The merged 'ABCDE' class for example contains 97% of the observed eyebrows movements and spans a time duration of 5183.79 seconds (about half of the whole corpus duration) splitted in 6469 intervals.

For comparison, we also ran our automatic annotation tool on the outputs of the OpenFace software for the 6 videos of the gold standard corpus. The results are illustrated figure 11. For class 'A' and 'B', the measures of precision are clearly worse when pipelining the OpenFace output. Whereas the overall coverage is comparable for the two outputs (for the merged class 'ABCDE', one obtains a recall of 0.96% for OpenFace versus a value of 0.97% for IntraFace), the IntraFace software output gives better results and has therefore to be preferred[1], at least for treating video corpus similar to our gold standard.

4. DISCUSSION

In this paper, we investigated the potential of two current state-of-the-art facial behavior analysis softwares (OpenFace and IntraFace) when applied to in-the-wild video corpus recording spontanenous facial expressions. For this purpose, the concrete case of eyebrows raising and frowning detection was considered. We proposed a tool which is pipelined with the output of the OpenFace or IntraFace software in order to produce an automatic annotation of eyebrows movements. The performance of our tool was hereafter evaluated by using our manually annotated corpus consisting of 573 eyebrow movements on 2h 50mn of video extracts as a gold standard. The results show that the tool can be already used with benefits for automatic annotation purpose (59% of success rate for the 'A' class and a coverage of 97% by analysing the half of the data).

Moreover, our study reveals that there is a room for improvement at the various stages of the overall process. The first stage concerns the experimental setup and the recording conditions of the corpus. It is important at this point that precise recommendations be delivered concerning camera positions, illumination conditions, ... As facial behavior analysis softwares perform better in general when recorded faces are close to frontal pose, cameras would ideally require to be positionned for fulfilling such a condition. Another example of recommendation would be to avoid occlusion events caused by an object or a second participant crossing the camera view.

[1]To pipeline the OpenFace software ouput was not initially planned by the authors. It unfortunately turned out that, during the course of the present study, the IntraFace software distribution stopped (i.e. *"Intraface is no longer available to download nor supported"*, see http://www.humansensing.cs.cmu.edu/intraface/), compelling us to consider an alternative facial behavior analysis software output to pipeline with.

Table 2: Evaluation table for the automatic annotation tool on the gold standard manually annotated (IntraFace output).

class	Npred	Nobs	precision	recall	match	duration
W	-	573	-	-	-	10209.89
X	-	25	-	-	-	740.20
M	-	548	-	-	-	9469.69
A	171	102	0.59	0.18	0.83	121.98
B	271	100	0.36	0.18	0.80	260.56
C	643	138	0.21	0.25	0.80	636.01
D	1723	124	0.07	0.22	0.69	1612.91
E	3661	70	0.02	0.13	0.63	2552.33
N	NA	14	-	-	-	4285.90
AB	442	202	0.45	0.36	0.81	382.54
ABC	1085	340	0.31	0.62	0.81	1018.55
ABCD	2808	464	0.16	0.84	0.78	2631.46
ABCDE	6469	534	0.08	0.97	0.76	5183.79

Figure 11: Comparison of the results of our automatic annotation tool pipelined with the IntraFace or OpenFace outputs.

A second point concerns the performance of the algorithms implemented in the facial behavior analysis softwares. We have shown section 3.4 that OpenFace was not as good as IntraFace for detecting eyebrows movements on our in-the-wild video corpus. One explanation could be that the CLM-Z algorithm [3] implemented in OpenFace for tracking facial features is less efficient than the IntraFace SDM algorithm [22]. It looks like OpenFace sometimes misinterpret eyebrows raising action as fictitious small head nod[2]. The accuracy and the robustness of the tracking facial features algorithm is of crucial importance since any detection errors introduced at this stage will propagate to the rest of the analysis. The treatment of in-the-wild videos is the today challenge of the computer vision communauty [20].

5. REFERENCES

[1] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3-4):273–287, 2007.

[2] T. Baltruišaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *11th IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[3] T. Baltruišaitis, P. Robinson, and L.-P. Morency. 3D constrained local model for rigid and non-rigid facial tracking. In *CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[4] T. Baltruišaitis, P. Robinson, and L.-P. Morency. OpenFace: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[5] J. Bavelas, J. Gerwing, and S. Healing. *Hand and facial gestures in conversational interaction*. Academic Press, 2014.

[6] C. Cavé, I. Guaitella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser. About the relationship between eyebrow movements and fo variations. In *Spoken Language*, pages 2175–2178, 1996.

[7] N. Chovil. Social determinants of facial displays. *Journal of Nonverbal Behavior*, 15(6):141–154, 1991.

[8] J. F. Cohn and F. De la Torre. Automated face analysis for affective computing. In R. Calvo, S. D'Mello, J. Gratch, and A. Kappas, editors, *The Oxford Handbook of Affective Computing (2014)*, 2014.

[9] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn. IntraFace. In *11th IEEE International Conference and Workshops On Automatic Face and Gesture Recognition*, 2015.

[10] P. Ekman and W. V. Friesen. *Facial Action Coding System: Manual*. Number vol. 1 à 2. Consulting Psychologists Press, 1978.

[11] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System: Research nexus*. Network Research Information, Salt Lake City, UT, 2002.

[12] M. L. Flecha-Garcia. Eyebrow raises in dialogue and their relation to discourse structure utterance function and pitch accents in english. *Speech Communication*, 52(6):542–554, 2010.

[13] J. Gorisch, C. Astésano, E. G. Bard, B. Bigi, and L. Prévot. Aix Map Task corpus: The french multimodal corpus of task-oriented dialogue. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2649 – 2652, Reykjavik, Iceland, May 26-31 2014.

[14] J. Gorisch and L. Prévot. Aix-DVD, LPL, 2014. https://www.ortolang.fr/market/corpora/sldr000891.

[15] A. Goujon, R. Bertrand, and M. Tellier. Eyebrows in French talk-in-interaction. In *Gesture and Speech in Interaction - 4th edition (GESPIN 4)*, pages 125–130, Nantes, France, September 2015.

[16] S. Mallat. *A wavelet tour of signal processing (second edition)*. Academic Press, 1999.

[17] B. Martinez, M. Valstar, B. Jiang, and M. Pantic. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 2017.

[18] D. McNeill, E. T. Levy, and S. D. Duncan. Gesture in discourse. In D. Tannen, H. E. Hamilton, and D. Schiffrin, editors, *The Handbook of Discourse Analysis*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2015.

[19] L. Prevot, J. Gorisch, and S. Mukherjee. Annotation and classification of french feedback communicative functions. *The 29th Pacific Asia Conference on Language, Information and Computation*, 2015.

[20] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[21] H. Sloetjes and P. Wittenburg. Annotation by category - ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

[22] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, June 2013.

[2]This remark is provisional and will obviously require further investigations.