Towards Learning Nonverbal Identities from the Web: Automatically Identifying Visually Accentuated Words

AmirAli B. Zadeh, Kenji Sagae, and Louis Philippe Morency

Institute for Creative Technologies,
University of Southern California,
12015 E Waterfront Drive, Playa Vista, CA 90094-2546, USA
{zadeh,sagae,morency}@ict.usc.edu

Abstract. This paper presents a novel long-term idea to learn automatically from online multimedia content, such as videos from YouTube channels, a portfolio of nonverbal identities in the form of computational representation of prototypical gestures of a speaker. As a first step towards this vision, this paper presents proof-of-concept experiments to automatically identify visually accentuated words from a collection of online videos of the same person. The experimental results are promising with many accentuated words automatically identified and specific head motion patterns were associated with these words.

1 Introduction

Much progress has been done in recent years in the field of interactive virtual agents. One particular emphasis has been on automatically generating nonverbal behaviors to accompany spoken words of the virtual human. While earlier versions of these nonverbal behavior generators [1, 2, 3, 5] were mostly based on literature review and general observations, new data-driven approaches [4, 6, 7, 8, 9] have been proposed to learn from a corpus of interaction the nonverbal behaviors that are used by a specific person or that generalizes across all participants. Given the significant cost associated with acquiring and annotating such dataset, an important issue is the scalability of these approaches. How can we create a large-scale portfolio of these nonverbal behavior generators for different interaction styles and personalities?

In this paper, we propose a long-term idea of using online multimedia content to help with the issue of scalability and customization of current nonverbal behavior generators. Video hosting websites such as YouTube contain a large amount of videos and channels where people are expressing their opinions about different topics. Each of these speakers has different communicative styles and their behaviors are idiosyncratic.

As a first step towards this vision, we present a proof of concept experiment focusing on visually accentuated words (i.e., spoken words that often co-occur with visual motion or emphasis) and their associated head gestures. We propose a statistical approach to automatically identify accentuated words in a collection of online videos from the same

T. Bickmore et al. (Eds.): IVA 2014, LNAI 8637, pp. 496–503, 2014.

speaker. We also present an analysis of the head motion patterns associated with these emphasized words, studying the variability and idiosyncrasy of these gestures. Our main research hypothesis is that some of the words will follow motion distribution different from the overall distribution, enabling us to identify these visually-accentuated words. Following a brief review of related work, section 3 describes our long-term vision. Our experiments and results are presented in Section 4 and 5. We conclude with future directions in Section 6. We will discuss the related works in the next section.

2 Related Work

Our research builds upon the previous literature and research on human gesture analysis and virtual human animation. Some of the original work on this topic used rule-based systems designed on general observations of human gestures [1, 2, 3]. Lee and Marsella created during their video analysis created a list of nonverbal behavior generation rules and used it for virtual human animation [5]. BEAT system developed by Cassell et al. uses plain text as input and based on priority values and linguistic analysis of the input text generates meaningful gestures and facial animation [13]. By focusing on behavior rules that would best generalize over a normal population, these approaches enabled automatic nonverbal behavior generation for virtual human.

Co-articulation of vocal and facial dynamics have also been studied alongside the relation between prosody features and certain gestures or gesture classes. Brand used a co-articulation model of vocal and facial dynamics to create realistically speaking animation [12]. Busso et al. created a dataset of videos and calculated rigid head motion. They used Hidden Markov Models for each emotional category they defined and synthesized a virtual human based on different emotions and prosody features [7].

More related to our current work, Stone et al. used a dataset of audio and motion captured segments of full body motion to recreate a specific person's gestures [6]. They use a set of simple grammar rules and match the gestures with the communicative function of the utterance. Neff et al. used an annotated dataset of two TV shows and created gesture profiles for each performer and later used these profiles in animation synthesis [4]. Their annotation scheme was based on a predefined gesture set and the customization for a specific individual was performed by directly modifying the statistics used to map gestures to semantic tags.

In contrast with prior work, our long-term vision covers the customization of both the timing of visual gestures as well as their appearance and dynamic. To reach this goal, we propose a novel approach which takes advantage of online multimedia content to automatically identify idiosyncratic gestures and their mapping to the verbal and prosodic content. As the first step towards this goal, we propose computational analysis of the relationship between head motion and spoken words to automatically quantify the nonverbal behaviors of a specific speaker.

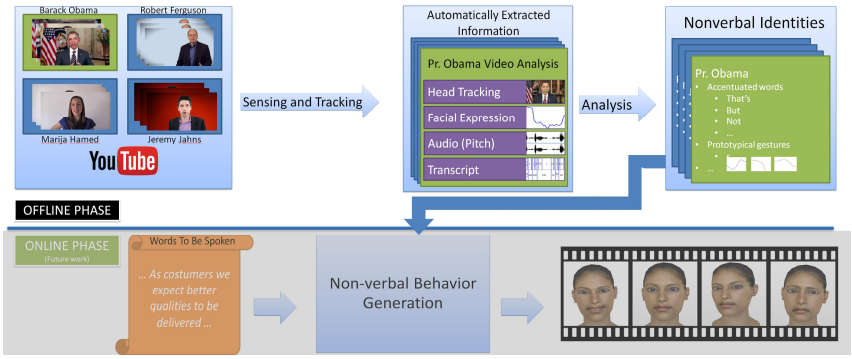

Fig. 1. An overview of our long-term vision for learning nonverbal identities from the web

3 Vision: Learning Nonverbal Identities from the Web

The long-term vision is to create a portfolio of *nonverbal identities* to help customize how virtual humans are animated. These nonverbal identities are computational representations of how a specific person gestures when speaking, including what words are usually emphasized, which gesture or facial expression is used to emphasize and how specific concepts or emotions are displayed by this speaker. To enable such large-scale diversity in computational representations of human nonverbal behaviors, we propose to take advantage of online multimedia content such as videos posted on YouTube to learn these nonverbal identities. These online websites are an almost infinite source of data since people love posting videos expressing their opinions about different topics.

Since many videos are often available for the same person (e.g., through a YouTube channel) it is possible to get multiple examples from the same person. With HD webcams becoming popular and microphone quality increasing, we can get high quality data on a large-scale. Figure 1 shows an overview of our long-term vision where nonverbal identities are learned automatically from online videos and then used to animate a virtual human that resembles a specific person (or a mixture of nonverbal identities). This vision includes an offline phase where online videos are analyzed to learn the portfolio of nonverbal identities and an online phase where these identities are used to animate the virtual human. The experiments presented in the following section focus on the offline phase, looking at the visually accentuated words.

4 Experiments: Identifying Visually-Accentuated Words

As a first step towards learning nonverbal identities from the web, we designed a preliminary set of experiments to study the viability of this approach, focusing on a specific type of nonverbal behavior (head motion) and its relationship to the verbal content. As discussed in the previous section, we are particularly interested in visually-accentuated

words and their related head gestures. Our primary goal with these experiments is to evaluate the feasibility of automatically extracting behavioral patterns from online videos (e.g., YouTube videos). As a secondary goal, we are interested to analyze the type of multimodal patterns identified in these online videos. In the following sub-sections, we present our approach for automatically crawling online videos, then present our techniques for automatic audio-visual feature extraction and finally describe our experimental methodology for our analysis.

Web Data Acquisition. Many online video sharing websites such as YouTube have a "channel" functionality where the same person (or company) can post multiple videos. These are particularly interesting in our case since we can easily gather multiple videos from the same person using these channels. In our experiments, we specifically used the White House Weekly Address video channel of President Barack Obama on YouTube. We developed a customized video web crawler which can find all videos of the same channel with all captions, audio and metadata of the online videos. This customized web crawler was augmented with a functionality to check if each download video contained only one person facing directly the camera. This functionality was optimized to work on a large-scale using CUDA and TBB. Our final dataset contained 196 videos of President Barack Obama, which represents more than 12 hours.

Audio-Visual Feature Extraction. An important aspect in our automatic multimodal feature extraction is the synchronization between information streams (text, audio and video). For the text modality, we took advantage of the captions associated to most YouTube videos. In fact, many videos posted on YouTube channels come with manually transcribed captions. Many companies are offering this service of manual transcription to YouTube video producers, simplifying our first step of speech transcription. To assure the synchronization of the text captions with the audio and video streams, we processed all videos using the P2FA forced alignment software [10]. This method allowed us to have exact timestamp for each spoken word. We assessed the quality of these word alignments on a subset of our dataset, showing good precision given the high quality recording of the audio stream.

Since we are interested in visual accentuation from the head motion, we automatically extracted head orientation from the video stream. For this step, we used the Intraface head pose tracker which returns a three dimensional rotation vector, representing the rotations around X, Y and Z axes [11]. These rotations can be interpreted as pitch, yaw and roll. These head orientation estimates were computed at 30Hz.

Methodology. The goal of our experiments is to study the interaction between spoken words and visual accentuations from head motions. To perform this analysis, we first created a dictionary of words spoken at least 50 times by President Obama. The stream of head orientation estimates was modified to compute the instantaneous head motion at each frame, keeping only the absolute value of this motion. This computation was performed for each rotation X, Y and Z. The multimodal analysis was performed by defining a time window of +/- 25 milliseconds around each instance of the words from our dictionary.

We are interested in automatically differentiating behavior patterns that are prototypical from the ones that are only happening by chance. As a first step in this direction, we propose to perform a statistical analysis to identify these recurring visual behaviors that are attached to specific verbal cues. To perform this analysis we hypothesize that the distribution of head motion happening during emphasized words is statistically different from the distribution of head motion over the whole interaction. To test this hypothesis, we perform student t-test analysis comparing all individual word with the overall distribution of head motion. This approach allows us to identify words with head motion patterns statistically different than the average spoken words. We can use the p-value returned by the statistical test as a measure of the uniqueness of this specific spoken word. This allowed us to identify the top visually accentuated words of President Obama for all three head rotations (pitch, yaw and roll). The following section describes our results and discussion.

Table 1. Top 10 words having lowest p-value in each rotation axis

	Rot. X	***Rot.Y***	***Rot. Z***
1	that's	not	but
2	but	all	that's
3	if	no	we
4	we	there's	I
5	I	the	it's
6	because	a	and
7	there's	across	why
8	so	just	so
9	don't	had	it
10	all	too	we're

5 Results and Discussion

Table 1 shows the top-ranked accentuated words by President Obama based on the head velocities around the X axis, the Y axis and the Z axis. The words presented in this table were ranked based on their p-value after the statistical test comparing them with the distribution of all words. A total of 155 words (including the words shown in Table 1) were shown to be statistically significant with $p<0.01$. This first result suggests that our multimodal analysis was able to automatically identify visually accentuated words. Another interesting result is the analysis of accentuated words per rotation axis. Words such as 'no' and 'not' are emerging for the rotation around Y axis which goes with our intuition that negative words should be accompanied by a head shake gesture. Words such as 'we' and 'I' are more significant around X and Z axes, which means either a gesture going vertically emphasizing himself (X axis) or tilting gesture including people around him. The number of significant words for the X, Y and Z rotations were 78, 85 and 91 respectively.

To better understand the head motion patterns accompanying these visually accentuated words, we plotted the average head rotational velocity around the X for 5 seconds before and after words from Table 1. Figure 2 shows these average graphs for three words: "don't", "because" and "but". We can see in all three cases an increase of the head motion around the word itself. It is important to notice that this head motion could be in either direction (e.g., going up or down for the rotation around the X axis).

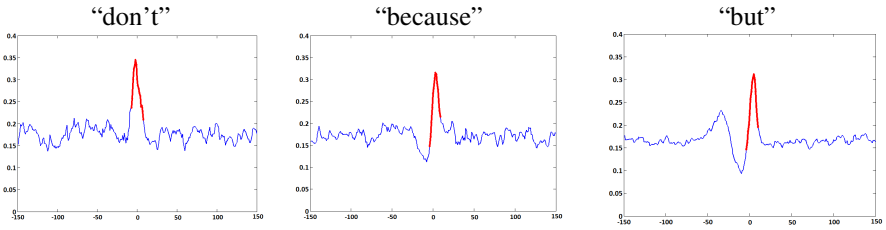

Fig. 2. Examples of the average motion plots (rotation around X axis, in degrees) for 3 top words from Table 1. The +/- 25 millisecond boundary of the specified word is drawn in red.

One interesting observation is the little dip right before the words “because” and “but”. This is most likely a preparation phase right before the emphasis of these two spoken words. Even more interesting is the second bump before the word “but” which means that the speaker also emphasized a previous word before not moving and then emphasizing the word “but”. During our analysis of the head motion patterns around this word “but”, we observed that a significant proportion of these instances are preceded by a short pause. In fact, this is a typical behavior of President Obama who often pauses for a little while before making his strong point, using a word such as “but”. These results show that our algorithms were able to identify such specific nonverbal behaviors of President Obama. This is a first step toward our long-term vision of automatically learning nonverbal identities of speakers based on their online videos.

We show in Figure 3 an example of a spoken sentence with below the direct head orientation around the X axis (i.e., pitch). We highlighted moments where the words “don’t’ were used. It is interesting to see that the head motion observed in Figure 2 is in fact a motion down for both instances. By segmenting these head motion instances we can start building a dictionary of prototypical head gestures and associate them with specific keywords. These behavioral rules can later be integrated in a generic nonverbal behavior generator to help customize it to a specific speaker.

6 Conclusion and Future Work

This paper introduced the long-term idea of learning prototypical nonverbal behaviors of a specific speaker from their online videos and use this information to customize the nonverbal behaviors of a virtual human. This automatic learning of *nonverbal identities* will allow us to create a portfolio of different speakers and enable more diversity in virtual human animations. As the first step towards this goal, we studied the relationship between head motion and spoken words to automatically identify visually accentuated words in online videos. Our results showed that we can automatically identify accentuated words for a specific speaker, showing interesting differences for head motion around the X, Y and Z axes.

... americans who still don't have jobs, but for the millions more who still don't have the right job ...

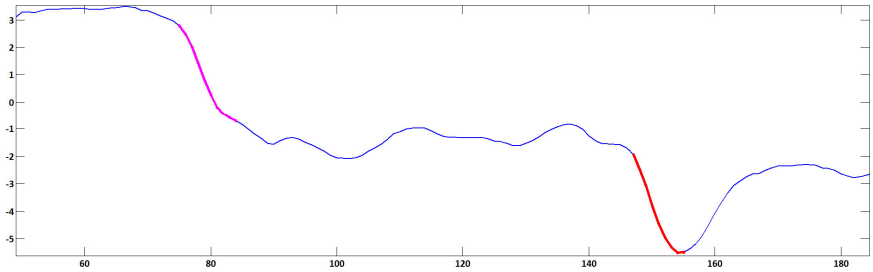

Fig. 3. Shows orientation around the X axis (pitch) for a specific spoken utterance. Two instances of the word "don't" are highlighted, showing in both case a motion downward

This first proof of concept opens up the way to many research directions analyzing online multimedia content to quantify human nonverbal behaviors. As one interesting next step, we plan to create a complete representation of not only the visually accentuated words but also include a dictionary of prototypical head gestures for each individual. We plan to evaluate the effectiveness of our virtual human animation method by studying how people perceive the virtual human gestures and if they are able to differentiate or even recognize specific person just from their customized virtual humans.

Acknowledgements. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1118018 and the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Douville, B., Prevost, S., And Stone, M.: Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents 413–420 (1994)
2. Decarlo, D., Stone, M., Revilla, C., And Venditti, J.J.: Specifying and animating fa-cial signals for discourse in embodied conversational agents. Computer Animation and Virtual Worlds 15(1), 27–38 (2004)
3. Bergmann, K., And Kopp, S.: Increasing the expressiveness of virtual agents: auto-nomous generation of speech and gesture for spatial description tasks. In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-, vol. 1, pp. 361–368 (2009)
4. Neff, M., Kipp, M., Albrecht, I., And Seidel, H.-P.: Gesture modeling and animation based on a probabilistic recreation of speaker style. ACM Transactions on Graphics 27(1), 5 (2008)

5. Lee, J., Marsella, S.C.: Nonverbal behavior generator for embodied conversational agents. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 243–255. Springer, Heidelberg (2006)
6. Stone, M., Decarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., Bregler, C.: Speaking with hands: Creating animated conversational characters from recordings of human performance. In: Proc. SIGGRAPH 2004, pp. 506–513 (2004)
7. Busso, C., Deng, Z., Grimm, M., Neumann, U., And Narayanan, S.: Rigid head motion in expressive speech animation: Analysis and synthesis. IEEE Transactions on Audio, Speech, and Language Processing 15(3), 1075–1086 (2007)
8. Albrecht, I., Haber, J., Peter Seidel, H.: Automatic generation of non-verbal facial expressions from speech. In: Proc. Computer Graphics International 2002, pp. 283–293 (2002)
9. Levine, S., Krähenbühl, P., Thrun, S., And Koltun, V.: Gesture controllers. In: ACM SIGGRAPH 2010 papers, SIGGRAPH 2010, pp. 124:1–124:11. ACM, New York (2010)
10. Yuan, J., Liberman, M.: Speaker identification on the SCOTUS corpus. In: Proceedings of Acoustics, pp. 5687–5690 (2008)
11. Xiong, X., De la Torre, F.: Supervised descent method and its applica-tions to face alignment. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2013)
12. Brand, M.: Voice puppetry. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1999, pp. 21–28. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA (1999)
13. Cassel, J., Vilhjálmsson, H., And Bickmore, T.: BEAT: The Behavior Expression Animation Toolkit. In: Proc. SIGGRAPH 2001, pp. 477–486 (2001)