ROC Comment: Automated Descriptive and Subjective Captioning of Behavioral Videos

Mohammad Rafayet Ali[1], Facundo Ciancio[2], Ru Zhao[3], Iftekhar Naim[1], Mohammed (Ehsan) Hoque[1]

Rochester Human-Computer Interaction (ROC HCI), University of Rochester, NY
{mali7, inaim, mehoque}@cs.rochester.edu[1], fciancio@ur.rochester.edu[2],
rzhao2@u.rochester.edu[3]

ABSTRACT

We present an automated interface, ROC Comment, for generating natural language comments on behavioral videos. We focus on the domain of public speaking, which many people consider their greatest fear. We collect a dataset of 196 public speaking videos from 49 individuals and gather 12,173 comments, generated by more than 500 independent human judges. We then train a k-Nearest-Neighbor (k-NN) based model by extracting prosodic (e.g., volume) and facial (e.g., smiles) features. Given a new video, we extract features and select the closest comments using k-NN model. We further filter the comments by clustering them using DBScan, and eliminating the outliers. Evaluation of our system with 30 participants conclude that while the generated comments are helpful, there is room for improvement in further personalizing them. Our model has been deployed online, allowing individuals to upload their videos and receive open-ended and interpretative comments. Our system is available at http://tinyurl.com/roccomment.

Author Keywords

Automated video captioning; objective feedback; comment generation; public speaking;

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

INTRODUCTION

Imagine that you would like to receive qualitative feedback on a speech you have prepared. One possibility is to record yourself and share the video with people you trust. That process, however, does not guarantee immediate feedback, and some might still feel uncomfortable sharing their video. What role could computing play in providing ubiquitous, automated, and immediate access to subjective and qualitative feedback on the recorded speech?

Motivated by the recent advances in automated image and video captioning, we explore the idea of automatically generating subjective comments for behavioral videos. Previous work on automated captioning aims to generate a natural language description of the objects and activities in an image/video [8,16]. Behavioral videos such as public speaking or job interview videos, however, have not been studied in this context. Generating comments about behavioral videos remains a difficult endeavor—primarily because it is not just about analyzing the pixels or attributes of the sequence of images, but also understanding how the dynamics of those differences can add to actionable recommendations or descriptions relevant to a real-world task (e.g., public speaking).

In this paper, we focus on the domain of public speaking, which is known to cause anxiety, fear, and even panic attacks [5]. Using our interface (ROC Comment), anyone can record a speech and receive interpretative comments on its quality—without having to share their videos with others.

To train our model, we have collected a dataset of 196 public speaking videos from 49 individuals with 12,173 comments from more than 500 human judges. Each human judge provided subjective comments on the public speaking skills demonstrated by the speaker. For each of the comments, we generate useful hashtags to summarize the comment. Our dataset is distinct that the videos are naturalistic, collected by allowing participants to record themselves in their environment using their laptop. Videos are given subjective and interpretative labels by independent judges, along with timing information regarding when those behaviors occurred in the video. Our system automatically extracts audio (e.g., volume) and video features (e.g., smile) from the training video, and aligns the features with human-generated comments to train a k-nearest-neighbor (k-NN)–based model. In the testing phase, the user uploads a video from which we extract facial and prosodic features. Using a time window, we then combine the features in the test video and select k (=10) windows in the training videos that have similar feature vectors. From these selected windows, we collect comments. To detect the irrelevant and too-specific

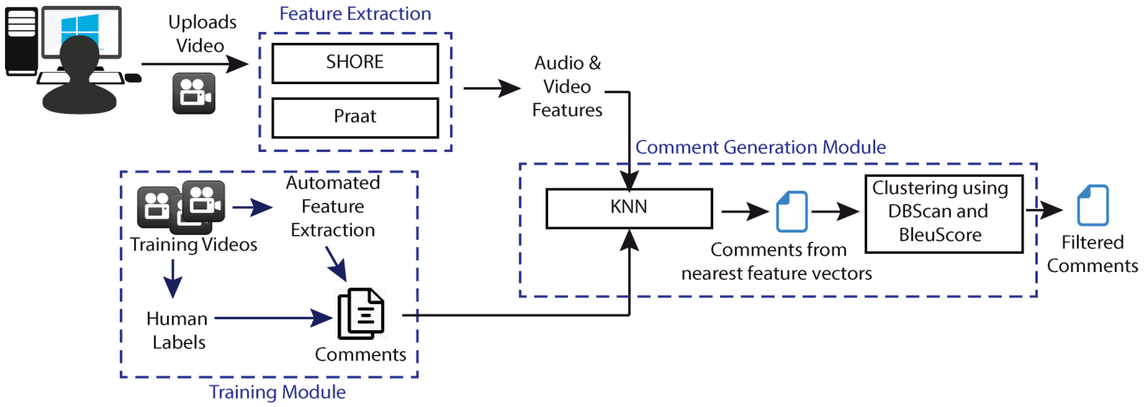


Figure 1: Using our interface, a user can record and upload his or her video. Our framework then automatically extracts the relevant audio and video features. Using the training data, our proposed k-NN–based model selects the nearest feature vectors and their corresponding comments. Using pairwise BLEU scores and DBScan, the outlier comments are identified and filtered from the output.

comments, we cluster the selected comments and find the outliers using density-based clustering(DBScan) [7]. We used BLEU [10] score as a proximity metric of the clustering algorithm. Thus, our developed interface can retrieve relevant comments for new test videos by analyzing the facial expressions and prosodic properties of the participants. We integrate our model with an end-to-end, fully automated, web-based user interface **(**http://tinyurl.com/roccomment**),** and allow individuals to upload and record their videos and receive open-ended, interpretative, and constructive comments, with hashtags summarizing their comments.

To validate ROC Comment, we ran a user study with 30 participants. Study results show that participants perceived the comments and the hashtags as helpful.

In this paper, we made the following contributions:

1. We proposed a new system for captioning behavioral videos in the context of public speaking.
2. We collected a new dataset of 196 videos of naturalistic public speaking experiences. We generated over 12,000 comments for training by recruiting 500 independent human judges.
3. We have developed a model utilizing k-NN to select comments similar to facial and prosodic features for a new video, and later filter the outlier comments using the DBScan clustering method.
4. Our algorithm has been instantiated and deployed online for anyone to try. We report results of a user study with 30 participants to inform our future work.

BACKGROUND

Over the last few years, there has been a growing interest in automated image captioning in both computer vision and natural language processing communities. The goal of automated image captioning is to generate seemingly human descriptions of an image. The existing image captioning methods can broadly be categorized into two groups: (1) k-nearest-neighbor (k-NN)–based approaches and (2) deep neural network–based approaches.

The k-NN–based models are conceptually simpler and have been shown to work well for image captioning [4,11]. These methods rely on a training dataset consisting of a large collection of images, each labeled with one or more human-generated captions. These methods extract k training images that are most similar to a test image and generate a new caption based on the human-generated captions of the training images. Further post-processing techniques have been applied to improve the captions' generalizability and relevance.

Deep neural network–based image captioning models jointly learn a neural language model for the captions in the training data and align different image regions with the corresponding words or phrases in the captions [10,17]. While these models work well in practice, they are more difficult to train, and require large amounts of training data to avoid overfitting.

Several recent papers have also studied the problem of automated video captioning. Given a short video clip, these methods generate a natural language sentence describing the objects and activities occurring in that clip. Some of the work include generating a semantic representation of the visual content via training a Conditional Random Field [12], using convolutional and recurrent neural networks [16], and using a dependency tree structure and deep neural network [18] to generate descriptions of short video clips.

In this paper, we describe how we apply k-NN to automatically generate interpretive and useful comments for behavioral videos, a previously unexplored area.

We apply our method to public speaking. To train for public speaking, both real-time and post feedback have been proven to be helpful. The Rhema [14] system uses Google Glass to give automated real-time feedback to a speaker as they speak. The TalkZones [13] system provides phone-based timing support while someone is speaking. ROC Comment is different than both, as it provides feedback after the speech.

The PitchPerfect [15] system helps improve public speaking by supporting structured rehearsal in slide-based presentation software. Our feedback mechanism is different, taking the form of interpretative comments. Later examples include the ROC Speak system, which utilizes crowds to generate feedback on speeches [9].

In this paper, we present ROC Comment that allows users to automatically receive subjective comments without having to share their videos with anyone else.

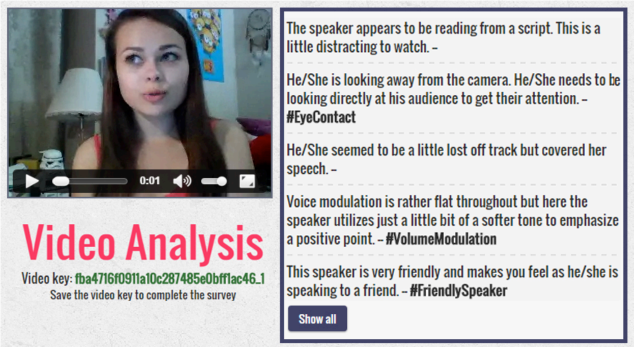
SYSTEM

We developed a web-based system where people can either upload their pre-recorded video or record their videos of public speaking and automatically receive comments. An example output is shown in Figure 2. Users can watch their video while reading the comments. Figure 1 shows the overall functionality of the ROC Comment system. We extracted the facial and prosodic features from the training videos and obtained comments with timestamps from human judges. The features are then aligned with comments using the timestamp. The comment generation module consists of a k-NN–based model and a cluster-based filtering method. From a test video, the k-NN–based model selects comments using the features of the video. We then used DBScan, which is a density-based clustering method, and find the outliers in the set of selected comments.

Dataset

Our dataset consists of 196 videos of 49 individuals giving speeches in front of their computers. The participants were recruited from Amazon Mechanical Turk. To achieve diversity in our corpus of recordings, we did not impose any constraint on recruitment. There were 22 female and 27 male speakers, with ages ranging from 20 to 60 years old. We gave the speakers a choice between five topics: a favorite hobby; how to find cheap airline tickets; how does real learning happen outside the classroom; whether children should watch less television; and a mock graduation speech. We asked them to speak in front of their web camera in a private space for approximately two minutes. To collect comments on the videos, we recruited raters from Amazon Mechanical Turk ("Turkers") and asked them to give at least three comments per video, with timestamp information, in three categories: body gestures, friendliness, and volume modulation. (One comment was required per category.) More than 500 Turkers (with a 95 percent acceptance rate) commented on the videos.

To generate hashtags, we took some sample comments and looked for keywords in those comments. Each keyword was associated with multiple hashtags. Then, for each comment in the training set, if it contains any keyword, a hashtag was assigned that was associated with it to reduce monotonicity.



Video key: fba4716f0911a10c287485e0bff1ac46_1
Save the video key to complete the survey

The speaker appears to be reading from a script. This is a little distracting to watch. –

He/She is looking away from the camera. He/She needs to be looking directly at his audience to get their attention. – #EyeContact

He/She seemed to be a little lost off track but covered her speech. –

Voice modulation is rather flat throughout but here the speaker utilizes just a little bit of a softer tone to emphasize a positive point. – #VolumeModulation

This speaker is very friendly and makes you feel as he/she is speaking to a friend. – #FriendlySpeaker

Figure 2: ROC Comment interface after receiving comments.

Challenges with Dataset

In the dataset, we faced several key challenges associated with real-world behavioral videos. The videos were recorded by 49 individuals, under different lighting and with different resolutions, which affected the extracted features. Different speakers kept different distances from their microphones, which resulted in volume variation within our dataset. Some people had cluttered backgrounds—for example, a picture of a face in the background—which added noise while tracking faces and detecting smiles. We normalized all extracted feature, which was, to some extent, able to address these problems.

Some inherent challenges came into play while collecting comments from the Turkers. Turkers who commented on the videos were not experts and had varied skill levels, backgrounds, and education. As a result, there was a large variation in the quality of the comments in our training set. Some comments were grammatically incorrect, less credible and less authoritative for end users. To filter these out, we used density-based clustering in the comments. Automatically eliminating grammatically incorrect sentences remains part of our future work.

Feature Extraction

We extracted both prosodic, and facial expression features from the training and test videos. We used the open-source speech analysis tool Praat [3] to extract prosodic features. The important prosodic features include pitch, vocal intensity, frequencies of the first three formants (F1, F2, F3), and average bandwidth. We extracted the smile intensity using the SHORE framework [19]. The value of the smile intensity was a positive integer between 0 and 100, where 0 indicates no smile and 100 indicates a full smile. We also extracted a measurement of body movement by estimating the images' pixel differences between consecutive frames. All these features were extracted to compose a 10-millisecond snapshot. For facial features, we took the average of the extracted features from the frames that lie between the 10-millisecond windows. To minimize the differences between multiple videos, we normalized all of the extracted features. For training and testing, we considered one-second-long segments and aggregated the features over the entire second by taking an average of all

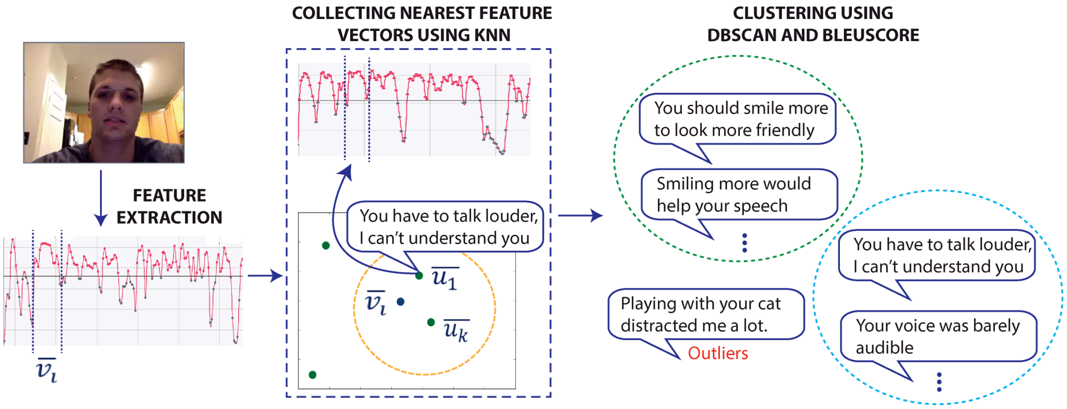

COLLECTING NEAREST FEATURE VECTORS USING KNN
CLUSTERING USING DBSCAN AND BLEUSCORE
FEATURE EXTRACTION
$\overline{v_i}$
You have to talk louder, I can't understand you
$\overline{u_1}$
$\overline{v_i}$
$\overline{u_k}$
You should smile more to look more friendly
Smiling more would help your speech
You have to talk louder, I can't understand you
Playing with your cat distracted me a lot.
Outliers
Your voice was barely audible

Figure 3: An example of comment generation from a user's video. From the uploaded video, features are extracted and feature vectors are created ($\overline{v}_i$). Using the k-NN model, the nearest feature vector that has comments associated with it are selected. From those comments, we perform clustering and identify the outliers.

10-millisecond windows in that segment.

Comment Generation Module

Our method incorporates ρ feature vectors by finding their average. From each of these newly-generated vectors, our model finds the k nearest vectors from the training set using Euclidean distance as a distance measure. As each comment in the training set has a timestamp, for each of these k neighbors, we find the comments which are not more than τ seconds apart. Then we output the unique comments. In this work, we set ρ = 100, τ = 5, and k = 10. The parameters were chosen by running this model on a small validation set of five videos and choosing the best one, based on human judgment. We eliminated gender-specific comments by replacing "he" and "she" with "he/she," and "his" and "her" with "his/her."

After selecting the comments from the nearest feature vectors, we calculate the pairwise BLEU score. Using the BLEU score as a proximity metric, we cluster the comments using density based clustering (DBScan). DBScan can discover clusters with arbitrary shape and unknown input parameters and label the outlier points, which cannot be assigned to any cluster. Using DBScan, we remove these outlier comments from the output. Figure 3 shows how the comments are being shown to users. Figure 4 shows the key points of our algorithm.

```
Input: $V$, feature vectors of test video
Output: comments
Procedure:
$i \leftarrow 0$
$S \leftarrow \emptyset$
While ($i < V.size$) do:
  $\vec{v} = avg(\vec{v}_i, \ldots, \vec{v}_{i+\rho})$
  $i \leftarrow i + \rho$
  Find $k$ nearest neighbor vectors $\vec{u}_1 \ldots \vec{u}_k$ from train set
  For each $\vec{u}_i \in \{\vec{u}_1 \ldots \vec{u}_k\}$ do:
    If a comment exists within $\tau$ sec of $\vec{u}_i$ do
      Add comment to set $S$
$D \in \mathbb{R}^{n \times n}$is a pairwise distance matrix, where $n = |S|$
For each $i, j \in S$ do:
  $distance[i][j] \leftarrow (1 - (Bleu(i,j) + Bleu(j,i))/2)$
Use DBScan to mark all outlier comments in $S$
$DBScan(S, distance)$
$Output \leftarrow i \in S \text{ and } i \text{ is not marked as Noise}$
```

Figure 4: Comment generation algorithm

EVALUATION

To evaluate ROC Comment, we ran a user study with n = 30 Turkers. In our guidelines, we provided a link to ROC Comment, directed participants to record a public speaking video approximately two minutes long, and asked 10 questions in an online survey. Our goal was to evaluate both the generated comments and system, overall. For this reason, we did not impose a time limit on preparing the speech before recording. Among the ten questions, seven were targeted to evaluate the usefulness, quality, and accuracy of the comments and hashtags. The other three asked whether users thought the comments came from a human or a computer algorithm, and why. In the first seven questions, we asked whether they agree or disagree with a statement, and the participants responded by giving a value from one to six, where one means strongly disagree and six means strongly agree. The statements and the average ratings (with standard deviation) are shown in Figure 5. Statement four and five were presented with an opposite sentiment to other statements to make the participants pay attention.

Participants thought that the comments were fairly helpful (avg. 3.53/6.00), and the hashtags were somewhat accurate (avg. 3.33/6.00). However, they agreed that "comments were not appropriate in the context of the speech" (avg. 3.90/6.00). Users found the comments somewhat out of context because, in our training data, some comments were context-specific. However, if we discard those participants who gave six (agree) for the "comments were not appropriate in the context of my speech" statement, we found that the average score of usefulness of the comments becomes 4.23. This indicates that, if we identify the context-specific comments and discard those, then the usability of the comments increases. Simple modification, such as, eliminating the comments, which contains the topic names, can reduce the problem to some extent. We found that there was indeed a negative correlation between the usefulness and out-of-context comments ratings (correlation = -0.81). There was also a high correlation between usefulness and accuracy (correlation = 0.83).

Our participants were unaware that a computer algorithm generated the comments. Five of our participants thought they came from real humans. Looking at their justification revealed informative insights. One participant said:

"The posture comment seems like it would not be computer generated as it would be something difficult for a computer to discern. However, the off topic comment makes me think it could be computer generated."

We noticed many relevant and helpful comments being generated by ROC Comment. From the generated comments, it is evident that some comments are indeed general on volume and friendliness attributes. As our model did not perform any language understanding, it sometimes picks up certain comments that do not match with the topic. This limitation was echoed by our participants as well:

"The comments felt a little generic to me and were not very accurate of the speech I gave. It is a very interesting new concept though."

FUTURE WORK

The simplicity and effectiveness of the k-NN–based models and cluster-based filtering motivated us to exploit them as a starting point. However, recent studies show that neural network–based captioning methods perform qualitatively better than k-NN–based models, despite achieving similar automated evaluation scores [6]. We plan to further expand our training dataset by including more video clips and human-generated comments, and apply neural network–based models to better understand the relationship between video features and corresponding natural language comments. Furthermore, we will consider the semantic features and sentiments of the comments in the training set, and exploit these signals for automatically composing novel comments by combining multiple relevant comments in the training data.

Our main goal was to generate comments that are specific enough to help people improve their public speaking skills and able to give insight. We plan to automatically detect the phrases that do not generalize well, and prune them using the tree-based model suggested in [4]. In the future, we plan to study behavioral videos beyond public speaking (e.g., job interviews, negotiations). So far, we only focused on providing comments on non-verbal behaviors (e.g., voice modulation, friendliness). Including verbal contents of the speech can be an exciting future endeavor.

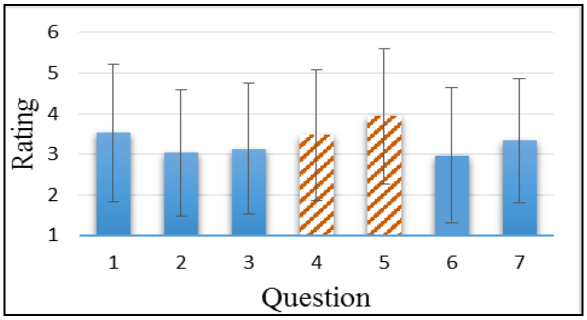


Figure 5: Average rating of seven statements: 1) the comments I received are very helpful; 2) the comments I received were accurate; 3) the comments I received were relevant; 4) the comments were too vague; 5) the comments were not appropriate in the context of my speech; 6) the hashtags were helpful; and 7) the hashtags were accurate. Here, ratings 1 = strongly disagree and 6 = strongly agree. Statements four and five were presented with the opposite sentiment.

Even though the smile intensity feature had a high correlation with the friendliness, we also noticed a few exceptions. Occasionally, we noticed speakers being perceived as friendly despite not smiling at all, as they expressed empathy and compassion through their spoken words and prosody.

The majority of the participants hinted that it is improbable for humans to provide that many comments so quickly, thereby, concluding that the comments are computer generated. It remains to be seen how the participants would have rated the system if we had introduced a delay in providing those comments instilling an illusion that those comments are coming from real people. Experimenting with this idea remains part of our future work.

CONCLUSION

We developed and deployed an online interface that allows users to either upload or record their speeches and automatically receive subjective comments. We developed our model using a k-NN model, and trained it on a new and naturalistic public speaking dataset, collected "in the wild." Generating automated and interpretive comments from behavioral videos has not been attempted in the past. In our initial exploratory work, we take on the challenge of collecting 196 naturalistic videos, and rigorously label them using online workers. We have developed a fully-automated online interface to determine the feasibility of our technique. While our algorithm could be improved further with more rigorous evaluation metrics, we feel that it is an exciting first step toward solving a difficult problem with immediate real-world implications.

ACKNOWLEDGEMENT

This work was supported by National Science Foundation Award #IIS-1464162 and in part by Grant W911NF-15-1-0542 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO).

REFERENCE

1. Nazia Ali and Ruchi Nagar. 2013. To study the effectiveness of occupational therapy intervention in the management of fear of public speaking in school going children aged between 12-17 years Methodology : 45, 3: 21–25.

2. E Boath, a Stewart, and a Carryer. 2012. Tapping for PEAS : Emotional Freedom Technique (EFT) in reducing Presentation Expression Anxiety Syndrome (PEAS) in University students . *Innovative Practice in Higher Education* 1, April: 1–12.

3. Paul Boersma and David Weenink. Praat: doing phonetics by computer. Retrieved from http://www.fon.hum.uva.nl/praat/

4. Yejin Choi, Tamara L Berg, U N C Chapel Hill, Chapel Hill, and Stony Brook. 2014. TREE TALK : Composition and Compression of Trees for Image Descriptions. 2: 351–362.

5. Purvinis Dalia and Susnienė Rūta. 2010. Insights on Problems of Public Speaking and Ways of Overcoming It. *Nation & Language: Modern Aspects of Socio-Linguistic Developmen;2010, p106.*

6. Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C Lawrence Zitnick. 2015. Exploring Nearest Neighbor Approaches for Image Captioning. *arXiv preprint arXiv:1505.04467*.

7. Martin Ester, Hans P Kriegel, Jorg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Second International Conference on Knowledge Discovery and Data Mining*: 226–231. http://doi.org/10.1.1.71.1980

8. Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, et al. 2010. Every picture tells a story: Generating sentences from images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6314 LNCS, PART 4: 15–29. http://doi.org/10.1007/978-3-642-15561-1_2

9. Michelle Fung, Yina Jin, Ru Zhao, and Mohammed Ehsan Hoque. 2015. ROC Speak: Semi-Automated Personalized Feedback on Nonverbal Behavior from Recorded Videos. *Proceedings of 17th International Conference on Ubiquitous Computing (Ubicomp)*.

10. Kishore Papineni, Salim Roukos, Todd Ward, and Wj Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Computational Linguistics (ACL),* July: 311–318. http://doi.org/10.3115/1073083.1073135

11. Polly Anne Rice. Emotional Freedom Techniques (EFT): Tap Into Empowerment. Retrieved from http://happyrealhealth.com/emotional-freedom-techniques-eft/

12. Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. 2013. Translating video content to natural language descriptions. *Proceedings of the IEEE International Conference on Computer Vision*, December: 433–440. http://doi.org/10.1109/ICCV.2013.61

13. Bahador Saket, Sijie Yang, Hong Tan, Koji Yatani, and Darren Edge. 2014. TalkZones: Section-based Time Support for Presentations. *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services (MobileHCI '14)*: 263–272. http://doi.org/10.1145/2628363.2628399

14. M Iftekhar Tanveer, Emy Lin, and Mohammed Ehsan Hoque. 2015. Rhema : A Real-Time In-Situ Intelligent Interface to Help People with Public Speaking. *IUI 2015: Proceedings of the 20th International Conference on Intelligent User Interfaces*, 286–295. http://doi.org/10.1145/2678025.2701386

15. Ha Trinh, Koji Yatani, and Darren Edge. 2014. PitchPerfect. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*: 1571–1580. http://doi.org/10.1145/2556288.2557286

16. Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2014. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

17. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and Tell: A Neural Image Caption Generator. Retrieved from http://arxiv.org/abs/1411.4555

18. R Xu, C Xiong, W Chen, and Jj Corso. 2015. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. *Proceedings of AAAI*. Retrieved from http://www.acsu.buffalo.edu/~rxu2/xu_corso_AAAI2015_v2t.pdf

19. SHORE^{TM} - Object and Face Recognition. Retrieved from http://www.iis.fraunhofer.de/en/ff/bsy/tech/bildanalyse/shore-gesichtsdetektion.html