Deep Multimodal Fusion for Persuasiveness Prediction

Behnaz Nojavanasghari[* 1] Deepak Gopinath[* 2] Jayanth Koushik[* 2]
Tadas Baltrušaitis[2] Louis-Philippe Morency[2]
[1] University of Central Florida, USA [2] Carnegie Mellon University, USA
behnaz@eecs.ucf.edu {dgopina1,jkoushik,tbaltrus,morency}@cs.cmu.edu

ABSTRACT

Persuasiveness is a high-level personality trait that quantifies the influence a speaker has on the beliefs, attitudes, intentions, motivations, and behavior of the audience. With social multimedia becoming an important channel in propagating ideas and opinions, analyzing persuasiveness is very important. In this work, we use the publicly available Persuasive Opinion Multimedia (POM) dataset to study persuasion. One of the challenges associated with this problem is the limited amount of annotated data. To tackle this challenge, we present a deep multimodal fusion architecture which is able to leverage complementary information from individual modalities for predicting persuasiveness. Our methods show significant improvement in performance over previous approaches.

CCS Concepts

•**Computing methodologies** → **Neural networks;** •**Applied computing** → *Psychology;*

Keywords

Persuasiveness, deep neural networks, multimodal fusion

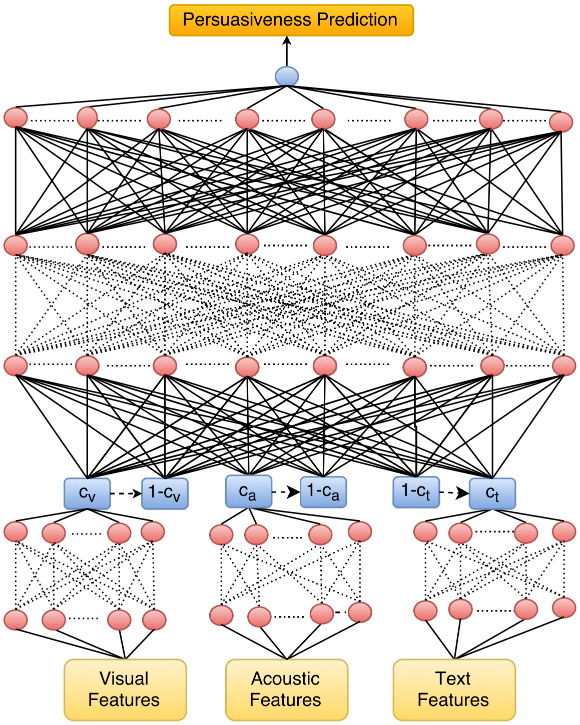

Figure 1: Architecture of our deep multimodal fusion model. c_v, c_a, and c_t represent the confidence scores of unimodal classifiers. These scores, along with the complementary scores ($1 - c_v$, $1 - c_a$, $1 - c_t$) are inputs to another deep neural network which makes the final prediction.

1. INTRODUCTION

With the advent of social networking websites and online collaboration tools, a lot of communication is happening online. Persuasive communication is the ability to influence one's beliefs, attitude, intentions, motivation, and behavior. Having the skill to be persuasive can be very useful in daily interactions, especially when the success of communication is dependent on being persuasive. For example, persuasion skills have a high impact on the performance of leaders [3]. Thus, a computational perspective on persuasion is very valuable, and can reveal the influential factors for persuasive communication. The findings of such studies could be useful for building automated training systems that provide feedback to people who desire to improve their persuasion skills.

Predicting persuasiveness is a challenging problem, as it depends not only on the words someone is uttering, but also on many other factors such as the visual behaviors that the person is displaying, and the way the person is uttering the words. This indicates that all modalities play a role in determining persuasiveness, and it is challenging to integrate information from multiple modalities.

[*] The indicated authors contributed equally to this work

ICMI'16, November 12–16, 2016, Tokyo, Japan
ACM. 978-1-4503-4556-9/16/11...$15.00
http://dx.doi.org/10.1145/2993148.2993176

Persuasiveness has been studied quite extensively from the psychological and social perspectives. The foundations and dynamics of persuasiveness, as well as theoretical frameworks on how to convey persuasive messages have been researched, and some of the most important factors, attitudes, and cognitive processes for determining persuasiveness have been reported [11, 21, 23]. However, there has been very little work on automatically predicting persuasiveness. The Persuasive Opinion Multimedia (POM) dataset is a multimodal dataset introduced by Park et al. [22] to study persuasiveness in social multimedia. There has been some research on this dataset; visual, acoustic, and verbal descriptors have been used to build unimodal and multimodal classifiers [7, 8, 22].

Siddiquie et al. [25] introduced the task of classifying politically persuasive web videos using the Rallying a Crowd (RAC) dataset [9]. The RAC dataset has videos where speakers are trying to persuade a crowd. Siddiquie et al. associate crowd reactions with persuasiveness, and use it as an extra cue in their predictions.

The previous works demonstrate the necessity of using all three modalities for predicting persuasiveness, and the advantage of multimodal approaches over unimodal ones. However, as the problem is challenging, there is a need for more complex architectures to learn and fuse the modalities.

Inspired by the success of deep learning techniques in various applications [26, 29], we present a deep multimodal fusion architecture for the task of persuasiveness prediction; our model has the ability to combine signals from the visual, acoustic, and text modalities effectively. Additionally, we address the problems associated with high dimensionality by using feature selection. To evaluate our proposed approach, we use the publicly available POM dataset (described in Section 4). Our method outperforms all prior work on the POM dataset confirming its effectiveness.

2. FEATURE DESCRIPTORS

Since automatic recognition of persuasiveness is not a trivial task, it is very important to identify and use the most important features for predicting it. We use high-level features to be able to identify and interpret the factors that have the most impact in differentiating persuasive videos from non-persuasive ones.

In the following sections, we describe the feature sets we use from each modality.

2.1 Visual Descriptors

Face movements and facial expressions have been identified as providing important information for interpreting emotional reactions and personality [14]. So we use the following visual features for our experiments[1]:

Presence and intensity of seven primary emotions and valence: Prior research shows that the presence of happiness, sadness, anger, and positive or negative attitudes can affect persuasiveness [19]. For this reason, we include as features the presence of seven primary emotions (anger, sadness, contempt, disgust, fear, joy, surprise), and the overall positivity or negativity of a video.

Activation of twenty elementary action units: Facial expressions can convey important signals about emotions, and influence judgment about persuasiveness [15]. Recognizing action units is important for facial expression analysis. We select the following action units as the second set of visual descriptors: inner brow raiser, outer brow raiser, brow lowerer, upper lid raiser, cheek raiser, lid tightener, nose wrinkler, upper lip raiser, lip corner puller, dimpler, lip corner depressor, chin raiser, lip puckerer, lip stretcher, lip tightener, lip pressor, lips part, jaw drop, lip suck.

Table 1: Some of the most discriminative features from the visual, acoustic, and text modalities. These were identified by the feature selection process.

Visual	Acoustic	Text
Upper lid raiser	MFCC 5	Highly
Surprise intensity	QOQ	Now
Presence of joy	F0	Laugh
Negativity intensity	Peak Slope	Amazing
Contempt intensity	NAQ	Absolutely

Head position and orientation: Head orientation and movements have been identified as being informative cues for recognizing persuasiveness [4]. So we use the movement, yaw, pitch, and roll of the head as the third set of visual features.

2.2 Acoustic Descriptors

Since prior research has shown that voice quality and prosody affect persuasiveness [1, 24], we use the following as our acoustic descriptors[2]:

Voice quality: Normalized amplitude quotient (NAQ), parabolic spectral parameter (PSP), maxima dispersion quotient (MDQ), quasi-open quotient (QOQ), difference between the first two harmonics (H1−H2), peak-slope, formants 1-5 (which show the creakiness, breathiness, and tenseness of speech signals).

Prosody: Pitch / fundamental frequency (F0) which measures the pitch or energy of voice.

MFCC: Standard Mel-frequency cepstral coefficients (MFCC 1–24)

2.3 Text Descriptors

To extract text descriptors, we treat transcript text as a bag of words. We build our vocabulary on the training set by retaining paraverbal features (like 'umm'), and removing all stop-words. Each transcript is represented as a sparse vector of tf-idf (term frequency-inverse document frequency) scores of vocabulary terms present in the transcript. This ensures that the weight of the terms is proportional to their frequency and specificity.

2.4 Feature Selection

Feature selection is the automatic selection of attributes from the feature set that are most relevant for building predictive models. In order to select features with the most predictive power, and to remove redundant features, we perform a t-test between the visual, acoustic, and verbal features extracted from persuasive and non-persuasive instances. We select features with p-values less than 0.05 [18]. This process is done using *only* the training set to increase the generalizability of the model on unseen data.

[1] Visual features were extracted using FACET: https://imotions.com/emotient/(accessed Sept-2016)

[2] Acoustic features were extracted using COVAREP [13].

2.5 Feature Analysis

Table 1 summarizes some of the most differentiating features for persuasiveness picked by the feature selection process. We analyze the selected features to see what type of features are identified as being important for persuasiveness.

Based on our analysis, surprise related behaviors such as the activation of upper lid raiser, and intensity of surprise were the top features selected from the visual modality. Expressing surprise can trigger thinking in the audience as it is unexpected. This leads the audience to think about the arguments that are being made by the speaker, and if they are strong, persuasion is significantly greater [16, 5].

From the acoustic modality, voice quality related features such as 'quasi-open quotient', 'peak slope', and 'normalized amplitude quotient' that identify the breathiness, tenseness, and pitch of the voice [12] were picked as discriminative features.

It is interesting to note that the words selected in this process such as 'highly', 'amazing', and 'absolutely' overlap with a list of persuasive words released by "The Father of Advertising", David Ogilvy in his book [20].

3. PROPOSED APPROACH

In this section we present details about our proposed unimodal (visual, acoustic, text), and multimodal approaches.

3.1 Unimodal Scheme

For training unimodal classifiers on the three modalities, we use a deep neural network as it allows the models to learn complex non-linear relationships between the input features. The loss function used is binary cross-entropy which is equivalent to the negative log likelihood, and is given by

$$L(\mathbf{x}, y) = -(y \log \sigma(\mathbf{x}) + (1-y)\log(1-\sigma(\mathbf{x}))) \quad (1)$$

Here $\mathbf{x}$ denotes an input feature vector from the training set, $y \in \{0, 1\}$ its true label, and $\sigma(\mathbf{x})$ indicates the predicted value (from the neural network). With this, the problem of finding the optimal set of parameters, θ^*, of the model can be formalized as

$$\theta^* = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} L(\mathbf{x}_i, y_i) \quad (2)$$

where $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ represents the training set with n training examples. This can be solved using stochastic gradient descent methods which, at the core, work by iteratively updating θ along the negative gradient of the loss. In our experiments, we use the Adam solver [17].

3.2 Multimodal Schemes

Persuasiveness is fundamentally multimodal in nature [22]. As mentioned in the introduction, being persuasive depends on the way someone conveys a message, which can be through visual, acoustic, and verbal signals. So, multimodal approaches are expected to perform better than unimodal approaches. In this section, we describe the early and late fusion approaches we explore in our experiments.

3.2.1 Early Fusion

For early fusion, we concatenate features from the three modalities into a single vector, and train a deep neural network on this new feature representation.

Table 2: Baseline results. This table compares the results of our baseline models with and without feature selection to majority voting, and previously reported results on the POM dataset[4].

Method	Accuracy	F_1 Score
Majority voting	0.62	0.76
SVM (all visual features)	0.59	0.62
SVM (selected visual features)	0.82	0.85
SVM (all acoustic features)	0.59	0.58
SVM (selected acoustic features)	0.72	0.77
SVM (all text features)	0.77	0.84
SVM (selected text features)	0.69	0.73
Park et al. [22]	0.71	-
Chatterjee et al. [8]	0.78	-

3.2.2 Late Fusion

While early fusion combines the different modalities in the initial phase, late fusion combines learned unimodal predictions into a final prediction. This leverages the individual potentials of unimodal classifiers. The deep neural networks trained for unimodal classification have a final single unit whose outputs can be interpreted as confidence scores for predicting the persuasiveness. In this work, we explore two schemes for combining these confidence scores in a late fusion framework.

Averaging: We average the confidence scores of individual unimodal classifiers to make the final prediction.

Deep Fusion: We use the final confidence score of each unimodal classifier (c), along with the complementary scores ($1 - c$) as input features to a fusing deep network. The intuition behind the addition of these complementary scores is that it helps the classifier infer the absence of persuasiveness. This architecture is illustrated in Figure 1.

4. EXPERIMENTS

Dataset: We use the Persuasive Opinion Multimedia (POM) dataset [22] introduced by Park et al. This dataset was collected with the goal of studying persuasiveness in a social media setting. We follow the experimental methodology proposed by the authors which leads to 130 persuasive videos, and 147 non-persuasive videos. Each instance is associated with the video and audio of a person (captured with a webcam), and a transcript of the spoken words.

Methodology: For our experiments, we split the dataset into training (205 videos), validation (33 videos) and test (39 videos) sets, ensuring that videos from the same person are not in two different sets. We extract features, as described in Section 2, from each modality. As the features of the visual and acoustic modalities are computed from short time windows, we use the mean, median, standard deviation, minimum, maximum, range (maximum - minimum), skewness, and percentiles (10th, 25th, 75th, and 90th) as a way of summarizing entire videos.

For the deep neural network architecture, we use a network with multiple fully connected layers, and add dropout [28] right after the input layer. We select the number of layers,

[4] It should be noted that Park et al. [22] and Chatterjee et al. [8] use a different testing methodology (n-fold).

Table 3: Results of unimodal classifiers. Note that feature selection improves the performance consistently across all modalities.

Modality	Accuracy		F_1 Score	
	All features	Selected features	All features	Selected features
Visual	0.54	**0.87**	0.57	**0.90**
Acoustic	0.74	**0.80**	0.77	**0.82**
Text	0.69	**0.85**	0.75	**0.88**

number of units in each layer, learning rate, amount of dropout, and number epochs by validating over a set of values. We vary the number of layers in {2, 3, 5, 7, 10}, the number of units per layer in {3, 5, 7, 10, 15, 20}, the learning rate in {0.0001, 0.0003, 0.001, 0.003, 0.01}, the dropout in {0.0, 0.1, 0.3, 0.5, 0.75, 0.9}, and the number of epochs in {300, 500, 800, 1000, 1200}. As there are many local optima, the models show high variance leading to suboptimal parameters being picked by cross-validation. To tackle this issue, we use an ensembling approach. During cross-validation, with each set of parameters, we train five times, and use average accuracy on the validation set to pick the best set of parameters. With this best set of parameters, we train 100 times on the training set, and select the models that perform best on the validation set. Using these final predictive models, we obtain results on the test set. We evaluate models using accuracy and F_1 score.

We developed our models primarily using the Keras library [10] for Python. The code, and dataset splits used for our experiments are publicly available[5].

5. RESULTS AND DISCUSSION

In this section, we discuss baseline models, compare classifiers with and without feature selection, and compare the performance of different fusion schemes.

Baseline models: We use support vector machines (SVM) to train our baseline models for all three modalities (see Table 2). We train these models on both the full feature sets, and reduced feature sets obtained after feature selection. We use the C-SVC [2] implementation provided by LIBSVM [6]. The hyper parameters of the SVM (C and γ) are selected from 100 uniformly spaced values in $[10^{-7}, 10^{7}]$ using cross-validation. We compare our baselines with the majority voting classifier, and the results of prior work on the POM dataset by Park et al. [22] and Chatterjee et al. [8].

Unimodal approaches: Table 3 summarizes the results of training unimodal classifiers on both the full set of features, and the reduced feature set obtained after feature selection. The performance of models after feature selection improves for all three modalities, and surpasses the results of SVMs. Based on these results, we make some interesting observations about the unimodal classifiers. For each modality, the best results are obtained by training a deep neural network on a reduced set of features. In a comparison between the performance of the unimodal classifiers, we found that despite previous research which has identified text to be the most important modality for persuasiveness [22], our best results are from the visual modality.

Table 4: Multimodal fusion results. The best results are achieved by using the proposed deep fusion framework.

Fusion Method	Accuracy	F_1 Score
Early Fusion	0.85	0.87
Late Fusion with Averaging	0.87	0.89
Late Fusion with DNN	**0.90**	**0.91**

Feature selection allows us to identify the attributes most relevant to persuasiveness (see Table 1), and remove unnecessary and irrelevant features that do not contribute to the accuracy of the predictive model. We believe that the initial set of features is noisy, and the model is not able to deal with the noisy training data. Lower variance in the feature set makes the models generalize better. Since the reduced feature sets produce the best results, we use these features, and the corresponding classifiers for our fusion models.

Multimodal approaches: Table 4 shows the performance of our multimodal approaches. As you can see, early fusion does better than the text and acoustic unimodal classifiers, but it does not perform as well as the other two fusion techniques. As we concatenate the features at an early stage, it increases the dimensionality, but not all features are very important for prediction. This can cause a lower accuracy for early fusion compared to late fusion.

For late fusion, we perform two sets of experiments: (1) averaging the confidence values from the three modalities, and (2) training a deep neural network which takes the output confidence scores (and complementary confidence scores) of the unimodal models, and makes the final predictions. The first method performs better than early fusion; and the latter outperforms both early fusion, and the averaging scheme. The reason for improved performance of late fusion compared to early fusion is that in early fusion, we concatenate all features and do not consider the fact that these features have different representations; we treat them equally. This can lead to lower performance of the classifier [27]. In late fusion with averaging, we again give the same importance to all three modalities, and if one modality is noisy, it can affect the final predictions. However, in late fusion with a deep neural network, the model learns the importance of each modality and is allowed to learn a non-linear combination of predictions and the weights contributing to the final model are assigned accordingly.

6. CONCLUSIONS

In this paper, we studied persuasiveness from a computational perspective and introduced a deep neural network architecture for predicting persuasiveness using the visual, acoustic, and text modalities. We showed that the proposed architecture is able to deal with limited labeled data while taking advantage of the ability of deep models in discovering complex relationships between input features. We developed a deep multimodal fusion model which improved the performance over unimodal models. Our experiments showed that all three modalities – visual, acoustic, and text can work complementary to each other for predicting persuasiveness. Finally, we demonstrated the utility of deep neural networks for performing late fusion by outperforming all previously reported methods on the POM dataset.

[5] https://github.com/jayanthkoushik/cmu-ammml-project

7. REFERENCES

[1] W. Apple, L. A. Streeter, and R. M. Krauss. Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, 37(5):715, 1979.

[2] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[3] R. Boyatzis, R. E. Boyatzis, and F. Ratti. Emotional, social and cognitive intelligence competencies distinguishing effective italian managers and leaders in a private company and cooperatives. *Journal of Management Development*, 28(9):821–838, 2009.

[4] P. Briñol and R. E. Petty. Overt head movements and persuasion: a self-validation analysis. *Journal of personality and social psychology*, 84(6):1123, 2003.

[5] J. K. Burgoon and S. B. Jones. Toward a theory of personal space expectations and their violations. *Human Communication Research*, 2(2):131–146, 1976.

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[7] M. Chatterjee, S. Park, L. P. Morency, and S. Scherer. Combining two perspectives on classifying multimodal data for recognizing speaker traits. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 7–14. ACM, 2015.

[8] M. Chatterjee, S. Park, H. S. Shim, K. Sagae, and L. P. Morency. Verbal behaviors and persuasiveness in online multimedia content. *SocialNLP 2014*, page 50, 2014.

[9] D. Chisholm, B. Siddiquie, A. Divakaran, and E. Shriberg. Audio-based affect detection in web videos. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.

[10] F. Chollet. Keras. https://github.com/fchollet/keras, 2015.

[11] W. D. Crano and R. Prislin. Attitudes and persuasion. *Annu. Rev. Psychol.*, 57:345–374, 2006.

[12] A. Cullen, J. Kane, T. Drugman, and N. Harte. Creaky voice and the classification of affect. *Proceedings of WASSS, Grenoble, France*, 2013.

[13] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. COVAREP- a collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 960–964. IEEE, 2014.

[14] U. Hess, R. B. Adams, and R. E. Kleck. The face is not an empty canvas: how facial expressions interact with facial appearance. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1535):3497–3504, 2009.

[15] J. Hyde, E. J. Carter, S. Kiesler, and J. K. Hodgins. Using an interactive avatar's facial expressiveness to increase persuasiveness and socialness. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1719–1728. ACM, 2015.

[16] U. R. Karmarkar and Z. L. Tormala. Believe me, i have no idea what i'm talking about: The effects of source certainty on consumer involvement and persuasion. *Journal of Consumer Research*, 36(6):1033–1049, 2010.

[17] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] H. Liu, J. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome informatics*, 13:51–60, 2002.

[19] M. Mitchell, K. Brown, M. Morris-Villagran, and P. Villagran. The effects of anger, sadness and happiness on persuasive message processing: A test of the negative state relief model. *Communication Monographs*, 68(4):347–359, 2001.

[20] D. Ogilvy and R. Atherton. *Confessions of an advertising man.* Atheneum New York, 1963.

[21] D. J. O'Keefe. *Persuasion: Theory and research.* Sage Publications, 2015.

[22] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L. P. Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 50–57. ACM, 2014.

[23] R. M. Perloff. *The dynamics of persuasion: communication and attitudes in the twenty-first century.* Routledge, 2010.

[24] K. R. Scherer, H. London, and J. J. Wolf. The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality*, 7(1):31–44, 1973.

[25] B. Siddiquie, D. Chisholm, and A. Divakaran. Exploiting multimodal affect and semantics to identify politically persuasive web videos. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 203–210. ACM, 2015.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 399–402, New York, NY, USA, 2005. ACM.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[29] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller. Deep neural networks for acoustic emotion recognition: raising the benchmarks. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011.