Multichannel attention network for analyzing visual behavior in public speaking

Rahul Sharma
IIT Kanpur
ra.rahulsharma.sh@gmail.com

Tanaya Guha
IIT Kanpur
tanaya@iitk.ac.in

Gaurav Sharma
IIT Kanpur NEC Labs America[1]
grv@cse.iitk.ac.in

Abstract

We investigate the importance of human centered visual cues for predicting the popularity of a public lecture. We construct a large database of more than 1800 TED talk videos and leverage the corresponding (online) viewers' ratings from YouTube for a measure of popularity of the TED talks. Visual cues related to facial and physical appearance, facial expressions, and pose variations are learned using convolutional neural networks (CNN) connected to an attention-based long short-term memory (LSTM) network to predict the video popularity. The proposed overall network is end-to-end-trainable, and achieves state-of-the-art prediction accuracy indicating that the visual cues alone contain highly predictive information about the popularity of a talk. We also demonstrate qualitatively that the network learns a human-like attention mechanism, which is particularly useful for interpretability, i.e. how attention varies with time, and across different visual cues as a function of their relative importance.

1. Introduction

Analysis and modeling of human behavior are critical for human-centric systems to predict the outcome of social interactions, and to improve interactions between humans or between human and computer. Human behavior is expressed and perceived in terms of verbal (e.g. spoken dialogs, pitch) and visual cues (e.g. hand and body gestures, facial expressions) [27]. These behavioral cues can be captured and processed to predict the outcome of social interactions. Public speaking is an important aspect of human communication. A good speaker is articulate, has convincing body language, and often, can significantly influence people [29]. While the success of public speaking largely depends on the content of the talk, and the speaker's verbal behavior, non-verbal (visual) cues such as gestures and physical appearance also play a significant role [19]. In this paper, we investigate the importance of visual cues for predicting popularity of speaker in public speaking videos.

[1]Gaurav Sharma is currently with NEC Labs America, majority of the work was done when he was with IIT Kanpur.

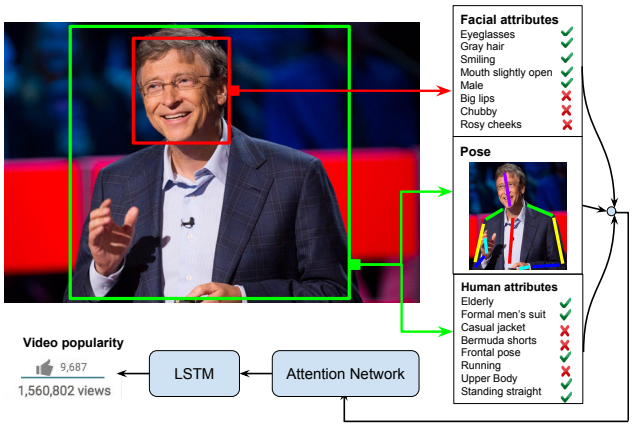

Figure 1. The main idea of the proposed framework for predicting public lecture popularity from visual cues.

We analyze videos from the very popular Technology, Entertainment, Design (TED) seminar series, and construct a database of 1864 TED talk videos, with associated statistics (number of *views, likes, dislikes* and *comments*) from YouTube. As a quantitative estimate of the popularity of a TED talk, we use the ratio of the number of *likes* to that of the *views*. We refer to this ratio as the *video popularity*.

We develop a computational framework for predicting the popularity of public speaking videos from visual cues. The main idea of our framework is summarized in Fig. 1. We hypothesize that the visual cues related to face, gesture, and physical appearance of a speaker together contribute to the popularity of a public lecture. This information is captured using three convolutional neural network (CNN) streams corresponding to different visual cues pertaining to physical appearance, gestures and facial expressions. These channels are then fused to predict video popularity. Motivated by the success of the long short-term memory (LSTM) networks in sequence prediction tasks [2, 33], our first approach to multichannel fusion is to concatenate the channel encodings (corresponding to the individual cues) to form a single monolithic feature. This serves as a competitive *baseline* for our task. The simple concatenation of the channel encodings however is sub-optimal, since the fea-

tures may lie in distinct spaces, with their respective different properties. To address this issue, we perform an alignment of the channels within the LSTM, and learn the alignment parameters along with all other parameters of the network. Further, we integrate an attention mechanism into the framework as follows. At every time step, we predict which CNN stream is the most relevant by learning the attention scores as a latent variable in the network. Incorporating multichannel attention gives us the benefit of interpretability, as the attention scores on the different cues provide insights to the relative importance of the visual cues over time.

The contributions of this work are as follows: (i) A large database containing videos of TED talks and their corresponding YouTube metadata is constructed to facilitate the study of public speaking in general, (ii) A novel architecture based on channel alignment and multichannel attention LSTM is proposed for predicting video popularity from visual cues. The network fosters interpretability and analysis of the visual cues, providing insight to the significance of different visual cues in public speaking.

2. Related work

Human behavior in the context of public speaking has been studied extensively from the psychological and social perspectives. For example, psychological studies have investigated the influence of non-verbal cues [19], importance of using confidence cues (phrases that express speaker's confidence) in speech [29], the effect of physical distance [1] in speaker's likability and persuasiveness, and the fear of public speaking [4, 18]. In the computational front however, work on automatic modeling, analysis and prediction of public speaking behavior is relatively limited.

The existing literature of computational analysis of public speaking suggests the use of various modalities including speech, video, motion capture (MoCap), and even, manual annotation of behavioral activities. The majority of work involves analysis of speech and verbal behavior. Related work on speech include acoustic, prosodic and lexical analysis to discover vocal characteristics of a good speaker [20, 25], and quantifying a speaker's attractiveness [8]. A shift-invariant dictionary learning method was proposed to detect human-interpretable behavioral cues such as hand gestures, pose, and body movements from MoCap data [26]. A database of political speeches along with perceptual ratings was constructed and used to study the role of vocal variety, voice quality, speech fluency, and pause timings on perceived speaking performance [20, 25]. Motion energy was also extracted as a visual cue, and was shown to have positive correlation with the manual ratings obtained for speaker performance. Acoustic characteristics of speech was also studied to quantify attractiveness and pleasantness of a speaker. The importance of these behavioral cues was studied in the context of public speaking using a database containing videos and MoCap sequences of 55 public speeches [26].

In another recent work, researchers attempted to automatically identify the nonverbal/visual behavioral cues that are correlated with human experts' opinion of speaker performance [30, 31]. An automatic performance evaluation was done using a database of 47 people presenting in front of a virtual audience. A related study on public speaking anxiety was also performed on the same database [31]. In a related work on job interviews, non-verbal behavioral cues were used to estimate a candidate's hirability [16]. More recently, a deep multimodal fusion architecture was proposed to predict persuasiveness of a speaker that indicates the influence a speaker has on the beliefs of an audience [17]. This framework used video, audio and text descriptors to predict persuasiveness on a publicly available database containing more than 200 videos. The descriptors used in their work consisted of standard acoustic and text features, and several hand crafted visual features.

Compared to the existing literature, the work presented in the current paper studies public speaking at a much larger scale, and in particular focuses on the visual aspects of public speaking. The approach is completely data-driven, and does not use any manual annotation for encoding behavioral cues. It uses the highly successful CNN architectures, namely the AlexNet [11], and the VGGNet [24] for capturing visual cues. The modeling of sequential data is based on a variant of the recurrent neural networks (RNN), called the LSTM [9]. The LSTM networks have been highly successful in addressing several visual and multimodal tasks, such as action classification [12], image captioning [32], and visual question answering [15]. The attention-based framework proposed in our paper is inspired by the success of the attention models used in various visual recognition tasks [10, 33]. Perhaps the attention network most related to ours is the stacked attention network (SAN) [33]. However, our method differs from SAN in the following ways: (i) SAN considers *spatial* attention for the task of visual question answering, while our network computes *temporal* attention across multiple channels for video popularity prediction (ii) SAN (and other attention networks) assumes the availability of information from the entire data to compute attention, while in our proposed architecture, attention is predicted at every time step, based on the information from the multiple channels in past frames, and those in the current frame.

3. Database creation

To facilitate the study of public speaking behavior, we constructed a large video database, namely the **TED1.8K** database. This database contains 1864 TED talk videos, and their associated metadata collected from YouTube. The YouTube metadata that we collected for each video are -

Figure 2. Sample frames from our TED1.8K database along with YouTube metadata.

Table 1. Overview of the TED1.8K database

Total videos	1864
Average duration	13.7 min
YouTube metadata (mean, range)	
Views	247K, (40 – 10264K)
Likes	3075, (0 – 113K)
Dislikes	174, (0 – 5750)
Comments	462, (0 – 26K)

Figure 3. Sample frames from some of the discarded videos

number of *likes*, *dislikes*, *views* and *comments*. Fig. 2 shows sample frames from our TED1.8K database, which demonstrates the huge variability, and the challenging nature of the database. Table 1 presents a summary of the database.

We first collected all the TED talk videos published until June 2016. We discarded the videos for which the YouTube 'views' field was empty, or speaker's body/face could not be detected in the majority of frames (e.g. talks accompanied with dance performance). After this screening, 1864 videos remained (see Fig. 3 for examples). We choose to estimate the popularity of the public lecture, y, as the ratio of its number of *likes* to that of the *views*. These scores are normalized and mean-centered before feeding them to the regression network. The TED1.8K corpus provides certain advantages for studying human behavior in public speaking. Firstly, the video content is carefully created to have well defined audiovisual structure, with subtitles and transcripts of the talks. Therefore, the database offers opportunities for rich multimodal studies. Secondly, the TED talks are of diverse topics, and popular worldwide. Hence, the YouTube ratings are expected to come from viewers with varied demographics, age group, and social background, making the ratings rich and reliable.

4. Proposed framework

In this section, we develop the complete framework for predicting public speaking video popularity from visual cues. Our framework comprises three parts: (i) a collection of independent CNN streams that captures the visual cues, (ii) an attention network that selects the most interesting visual stream at every time step, and (iii) an LSTM network that predicts the popularity scores. Fig. 4 shows the complete architecture that we propose, and below, we describe each part in detail.

4.1. Encoding visual cues using CNN

Consider a video $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T]$, where $\mathbf{v}_i$ denotes the ith frame of the video. Each video $\mathbf{V}_j$ is associated with a corresponding video popularity score $y_j \in \mathbb{R}$ (see section 3). In order to visually describe a speaker's presence in a video frame, we extract the following visual cues: physical appearance, gestures and facial expressions. Given a video, we first detect the speaker at every frame by running a face detector [28] and a person detector [13]. Next, we set up three CNNs which encode the visual appearance and behavior of the speaker.

Facial attributes CNN: This network encodes facial attributes, such as smile, hairstyle, facial shape, and eye glasses (see Fig. 1). We expect such encoding of facial expressions and other attributes to contribute to the popularity of a public lecture. This CNN takes the primary face detected in $\mathbf{v}_i$ as an input, and encodes the facial attributes to a descriptor $\phi(\mathbf{v}_i)$.

Pose CNN: To account for the pose of a speaker, we use another CNN that predicts 17 landmark points in a human body, e.g. knees, elbows (see Fig. 1). This network encodes the body posture of a speaker in a frame. This feature describes the evolution of the body movements of the speaker as the talk goes on. We denote the corresponding descriptor as $\theta(\mathbf{v}_i)$ for frame $\mathbf{v}_i$.

Physical attributes CNN: Finally, a third CNN is used to encode the general full body attributes of a speaker, such as, gender, clothing, and age (see Fig. 1). As noted in psychological studies [19], the perceived impression of a speaker is likely to be influenced by such physical attributes, and hence we include them as a possible option to attend to by the full network. We denote the full human attribute descriptor as $\psi(\mathbf{v}_i)$ for frame $\mathbf{v}_i$.

4.2. Multichannel attention network

We build an attention network that systematically assigns weights to the three different descriptor channels at every

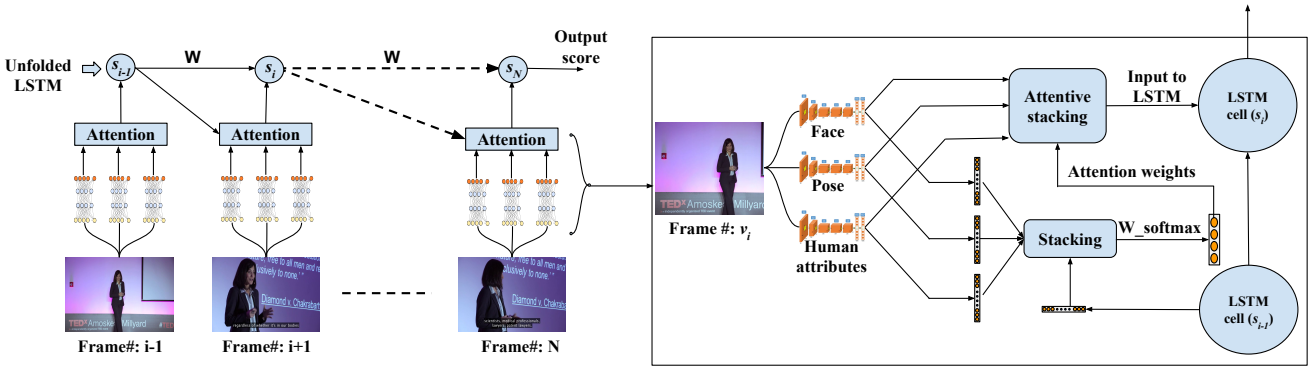

Figure 4. Complete architecture of the proposed attention-based LSTM framework.

time step. The computation of weights is based on the content of the current frame, as well as the information history from all the past frames.

Why attention? The intuition behind building the attention network is that viewers do not attend to all the visual cues simultaneously with equal importance. This is even more relevant in the current set up, where ratings from online viewers are being considered. It is possible that due to the variations in camera angle, and editing style, the upper body of a speaker is not properly visible or detected faces are too small (see Fig. 6). In such cases, the viewers are likely to rely on the visual cues that are easier to observe. On the other hand, due to occlusion, low illumination, camera angle, and other factors, the computation of the frame-level features can also introduce errors. Since our network relies on the history of all previous frames, an attention-based fusion can help in avoiding propagation of errors.

Given a frame $\mathbf{v} \in \mathbf{V}$, we obtain the three feature channels $\mathbf{f} = \phi(\mathbf{v})$, $\mathbf{p} = \theta(\mathbf{v})$ and $\mathbf{c} = \psi(\mathbf{v})$ corresponding to speaker's facial attributes, pose, and physical attributes by doing a forward pass of the respective CNNs. To align the different channels, we pass them through separate fully connected layers, which maps them to a common output space.

$$\begin{aligned} \mathbf{h}_\mathrm{f} &= \tanh(\mathbf{W}_\mathrm{f}\mathbf{f} + \mathbf{b}_\mathrm{f}) \\ \mathbf{h}_\mathrm{p} &= \tanh(\mathbf{W}_\mathrm{p}\mathbf{p} + \mathbf{b}_\mathrm{p}) \\ \mathbf{h}_\mathrm{c} &= \tanh(\mathbf{W}_\mathrm{c}\mathbf{c} + \mathbf{b}_\mathrm{c}) \end{aligned} \tag{1}$$

where, $\mathbf{W}_\mathrm{f} \in \mathbb{R}^{d_f \times m}, \mathbf{W}_\mathrm{p} \in \mathbb{R}^{d_p \times m}, \mathbf{W}_\mathrm{c} \in \mathbb{R}^{d_c \times m}$ are the projection weights of the three features' aligned representations, d_f, d_p, d_c are the feature dimensions respectively, m is the projected feature dimension, and $\mathbf{b}_\mathrm{f}, \mathbf{b}_\mathrm{p}, \mathbf{b}_\mathrm{c}$ are the corresponding biases. The information from all the past frames is encoded in $\mathbf{h}_\mathrm{s}$ as $\mathbf{h}_\mathrm{s} = \tanh(\mathbf{W}_\mathrm{s}\mathbf{s} + \mathbf{b}_\mathrm{s})$ where $\mathbf{s}$ and $\mathbf{W}_\mathrm{s} \in \mathbb{R}^{d_s \times m}$ denote the LSTM states, and the projection weights for the LSTM states, d_s being the dimension of the LSTM states. The aligned features, and the history from past frames are then passed through a 2-layer neural network with a softmax in the end to generate attention weight distribution over the three channels.

$$\begin{aligned} \mathbf{h}'_\mathrm{j} &= \tanh(\mathbf{W}_\mathrm{a}\mathbf{h}_\mathrm{j} + \mathbf{b}_\mathrm{a}) \;\; \forall j \in \{\mathrm{f}, \mathrm{p}, \mathrm{c}, \mathrm{s}\} \\ \mathbf{h}_\mathrm{a} &= [\mathbf{h}'_\mathrm{f}, \mathbf{h}'_\mathrm{p}, \mathbf{h}'_\mathrm{c}, \mathbf{h}'_\mathrm{s}] \end{aligned} \tag{2}$$

$$\mathbf{a} = \mathrm{softmax}(\mathbf{W}_\mathrm{sm}\mathbf{h}_\mathrm{a} + \mathbf{b}_\mathrm{sm}) \tag{3}$$

where $\mathbf{W}_\mathrm{a} \in \mathbb{R}^{m \times n}$ (and $\mathbf{W}_\mathrm{sm} \in \mathbb{R}^{4n \times 3}$ below) are the weights for the multichannel attention layer. Finally, $\mathbf{a}$ contains the attention weights corresponding to the facial attributes, pose and physical channels. Once the attention weights are computed, the channel having the maximum attention weight is selected. Let this channel be denoted as $\mathbf{h}^*$, where $\mathbf{h}^* \in \{\mathbf{h}_f, \mathbf{h}_p, \mathbf{h}_c\}$. Ignoring the other channels, only $\mathbf{h}^*$ is input to the LSTM network for regression (described below).

4.3. LSTM for regression

As the last part of the proposed architecture, we use LSTM [9] to model the video data as a sequence of frames. In our architecture, the LSTM cell takes an input $\mathbf{x}_t$ at every time step t, and updates the memory cell $\mathbf{s}_t$ in consultation with its previous state $\mathbf{s}_{t-1}$. At any time step t, $\mathbf{x}_t = \mathbf{h}^*_t$ where $\mathbf{h}^*_t$ is the feature vector with the highest attention at time instant t. The LSTM is equipped with several gates which control the update process. A forget gate $\mathbf{g}_t$ decides how much information from the past state $\mathbf{s}_{t-1}$ is carried forward. An input gate $\mathbf{i}_t$ supervises the information from the current input vector $\mathbf{x}_t$. An output gate $\mathbf{o}_t$ puts a check on the information that need to fed to the output as a hidden state. The state update process is as follows:

$$\begin{aligned} \mathbf{g}_t &= \sigma(\mathbf{W}_g\mathbf{x}_t + \mathbf{W}_g\mathbf{h}_{t-1} + \mathbf{b}_g) \\ \mathbf{i}_t &= \sigma(\mathbf{W}_i\mathbf{x}_t + \mathbf{W}_i + \mathbf{h}_{t-1} + \mathbf{b}_i) \\ \mathbf{s}_t &= \mathbf{g}_t\mathbf{s}_{t-1} + \mathbf{i}_t \tanh(\mathbf{W}_s\mathbf{x}_t + \mathbf{W}_s\mathbf{h}_{t-1} + \mathbf{b}_s) \\ \mathbf{o}_t &= \sigma(\mathbf{W}_o\mathbf{x}_t + \mathbf{W}_o\mathbf{h}_{t-1} + \mathbf{b}_o) \end{aligned} \tag{4}$$

$$\mathbf{h}_t = \mathbf{o}_t \tanh(\mathbf{s}_t) \tag{5}$$

where, the $\mathbf{W}$s and $\mathbf{b}$s are the weight matrices and the bias vectors that are learned during the training phase.

5. Performance evaluation

We perform extensive experiments to evaluate the proposed framework on the TED1.8K corpus (described in Sec. 3). Our experiments provide insights on how various visual cues contribute to the popularity of a public speaking video. We also analyze the system qualitatively by visualizing how attention is learned in our network.

5.1. Implementation details

The database is randomly divided with ratios $60:20:20$ to create the `train` , `val` and `test` sets. We report the mean of the performances on the test sets on multiple runs using different random splits. We plan to make the dataset, the splits and evaluation protocol public, upon publication. We consider two metrics for evaluating the performance of our systems: (i) the Pearson's correlation (ρ), and (ii) the mean square error (MSE). Both are computed between the predicted scores and those obtained from YouTube.

Facial attributes network: We use the AlexNet [11] for extracting the facial attribute features. The AlexNet is trained on the CelebA database [14] to perform facial attribute classification. CelebA is a large facial attributes database consisting a total of $200,000$ images, each with 40 binary attributes. The attributes are diverse, and cover many facial properties, such as 'smile', 'mouth slightly open', 'black hair', 'oval face' and 'mustache'. While training, we initialized the AlexNet with the model parameters pretrained on the ImageNet [21]. After training, we obtain a mean class classification accuracy of 79.27% on the CelebA database, which is comparable to the state-of-the-art results reported on the same database [14]. Next, we detect a speaker's face in every frame using a face detector [28]. If multiple faces are detected in a frame, we discard the frame due to the ambiguity in determining the speaker's face. We then feedforward the detected face through the trained AlexNet. The 4096-dimensional output from the last fully connected layer is our facial attributes descriptor.

Human body detection: To capture the pose of a speaker in a frame, we employ the single shot multibox detector (SSD) [13]. The SSD is an augmented version of VGGnet [24] trained on the VOC2007 database for general object detection, where one of the object categories is 'human'. We validated that the SSD can recognize the 'human' class with an accuracy of 72.5% on VOC2007 [5]. Using the SSD detector, we detect the speaker body at frame-level, and process the cropped part of the frame to obtain the pose and human attributes descriptors.

Pose network: A pretrained VGGnet [7] is used to obtain the pose descriptors. This VGGnet is originally trained for keypoint localization and action classification in unconstrained images. The pretrained VGGnet is validated on the VOC2012 database [6] that gives an accuracy of 70.5% over different action classes. To obtain the pose descriptor, we use the 4096-dimensional output from the last fully connected layer of this network.

Table 2. Performance of SVR in terms of correlation (ρ) for different scales and pooling strategies.

		Pooling operation			
	Scale	Mean	Stdev	Max	Grad
Face	single	0.27	0.30	0.29	0.27
	multi	0.30	0.42	**0.44**	0.39
Pose	single	0.19	0.24	0.28	0.22
	multi	0.24	0.29	**0.35**	0.27
Human attr.	single	0.27	0.30	0.35	0.28
	multi	0.36	0.40	**0.45**	0.32

Human attributes (HAT) network: Similar to the facial attributes features, we use the 4096 dimensional last fully connected layer output of the AlexNet to obtain the human attribute features. This AlexNet is trained on a human attributes database [23] with 9344 human images with 27 binary attributes. These attributes are related to the physical appearance of a human, such as, 'elderly', 'wearing t-shirt', 'female long skirt', and 'standing straight'. The trained model was validated on the human attributes database [23] to yield a mean average precision of 63.7% which is comparable to an earlier report [23].

Alignment and attention: We downsampled the videos at 5 frames per second to reduce the amount of data to be processed primarily due to the practical limitations on GPU memory size. We max-pool the descriptors (based on ablation experiments detailed below) within a volume of 11 frames (± 5 frames around the central frame) and a stride of 4. Thus, each volume contains information from ~ 2 seconds of a video. As described in Section 4.2, we align the descriptors by projecting them onto a common output subspace. We learn 3 representational layers, one for each descriptor, which projects the respective descriptor to a 1024 dimensional output. The aligned descriptors, and the LSTM states from the previous time step are compressed to 128 dimensions by another fully connected layer, and are stacked together (see Eq. (2)). Finally, they are feedforwarded through the last layer, and a *softmax* is applied to obtain attention distribution over different channels (see Eq. (3)). We use MSE as the loss function, and the RMSProp gradient descent method to learn the parameters of the LSTM and the attention network. Note that the proposed method is trainable end-to-end. However, we did not backpropagate the error into the visual cues networks while training the full system for predicting video popularity.

5.2. Baselines

We set up two baselines to compare with the proposed approach: (i) support vector regressor (SVR) that uses fixed length vectors as input, and (ii) LSTM (without alignment)

Table 3. Performance of LSTM (on a subset of TED1.8K) in terms of correlation (ρ) for frame-level and volume-level features.

Features	Frame-level	Vol-level multiple pool	Vol-level max pool
Face	0.45	0.51	**0.57**
Pose	0.31	0.39	**0.41**
Face + Pose	0.47	0.51	**0.58**

that uses time varying vector sequences as inputs. The details are provided below.

Support vector regression (SVR): To create a video-level feature from the frame-level features in a video, we use the pooled time series (PoT) representation [22] scheme. In this scheme, each dimension of a frame-level feature is considered as a time-series i.e. for d-dimensional frame-level features, there are d time-series. For each such time series, we perform 4 pooling operations i.e. mean pooling, standard deviation (stdev) pooling, max pooling, and histogram of time series gradient (grad) pooling. Each temporal filter pools over a window with varying size. A temporal pyramid structure [3] is created to pool temporal information at 5 different scales.

LSTM regression: We set up an LSTM (without alignment) with 50 hidden nodes for predicting the video popularity. This network uses a temporal max pooling within a video volume of 11 frames with a stride of 4 (similar to alignment and attention LSTM we propose).

5.3. Parameter settings

In this section, we discuss the effects and choices of different parameters in our framework.

Pooling operation: We run several prediction experiments using SVR with different pooling operations and temporal scales. Single scale indicates that the window size is equal to the video length, while multi-scale indicated the 5-level temporal pyramid scheme. The results in Table 2 indicate that multiscale max pooling consistently performs better than the other pooling strategies. Table 2 presents results for the individual channels for brevity. Similar trend was observed for the combinations of the channels as well. Hence, we choose *multiscale max pooling* for all our experiments with both SVR and LSTMs. We also experimented with LSTM-based prediction using features computed (i) at every frame, and (ii) at every volume consisting of 11 frames. For a given volume, the frame-level features are max pooled to compute the volume-level feature. Table 3 shows the performances of these two types of features, where volume-level features outperform the single frame-based ones.

Projection dimension for alignment: Recall that we proposed to align the individual channels to a common output (see Section 4.2). In order to do that the initial 4096-dimensional descriptors are projected onto a smaller dimension. Fig 5 shows how the prediction performance of each channel varies with different projection dimensions. Fig. 5 indicates that the 1024-dimensional representation performs the best for all channels except for human attributes. Since our architecture demands the same alignment dimension for all the channels, we chose to use 1024.

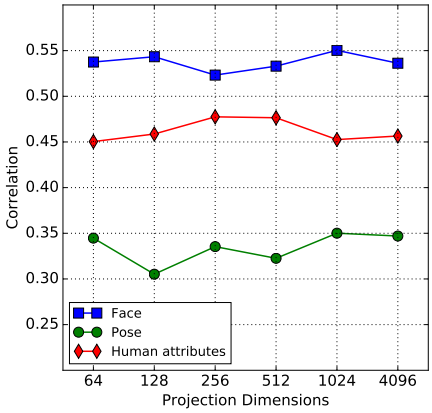

Figure 5. Performance variation of individual channel with projection dimension in alignment.

Table 4. Performance of the proposed framework on TED1.8K

Method	Pose HAT	Face HAT	Face Pose	Face,Pose HAT
	Correlation (ρ)			
SVR (baseline)	0.44	0.47	0.46	0.48
LSTM (baseline)	0.45	0.51	0.52	0.51
Aligned LSTM	**0.45**	0.51	0.52	0.51
Attn. LSTM	0.41	**0.51**	**0.55**	**0.57**
	MSE			
SVR (baseline)	0.87	0.84	0.87	0.76
LSTM (baseline)	0.81	0.72	0.71	0.73
Aligned LSTM	**0.74**	**0.70**	0.69	0.76
Attn. LSTM	0.80	0.74	**0.68**	**0.68**

5.4. Results

The performance of the proposed framework for predicting video popularity from visual cues is validated on the TED1.8K database. Our proposed aligned and attention-based LSTM networks are compared against the two challenging baselines, SVR and LSTM. Experiments have been carried out for all possible combinations of the visual channels. Table 4 summarizes the performances of the proposed LSTM networks and the baselines in terms of correlation ρ and MSE.

SVR vs. LSTMs: Table 4 shows that LSTMs (all variants) perform better than SVR in every case, i.e. for any combination of the visual channels. The performance improvement

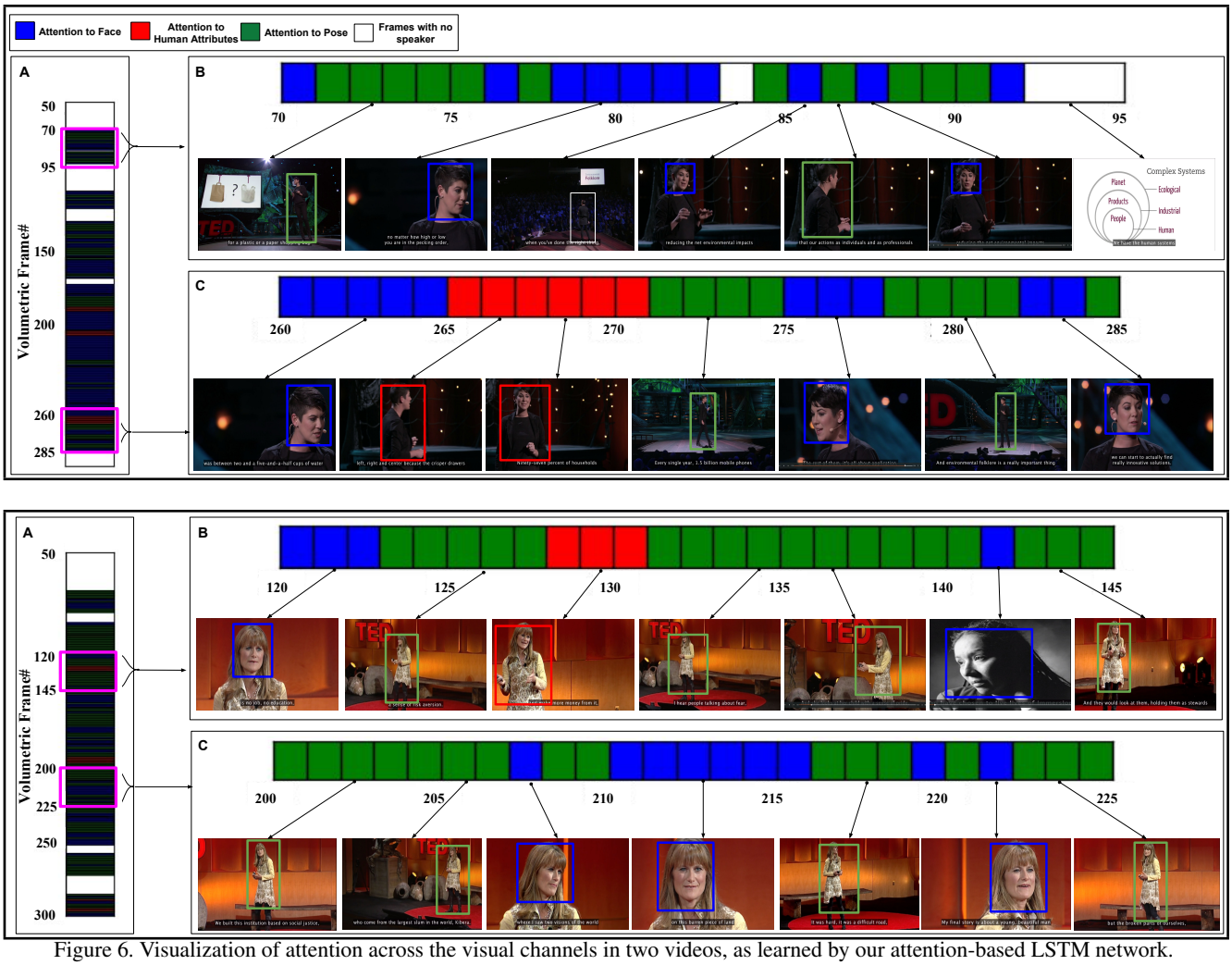

Figure 6. Visualization of attention across the visual channels in two videos, as learned by our attention-based LSTM network.

of LSTM is notable when facial attributes and pose channels are used in the network. A significant upgrade in ρ from 0.45 (SVR) to 0.52 is obtained for both the baseline LSTM and the aligned LSTM. Accordingly, a significant drop in MSE values from 0.87 to 0.71 is observed for the case of facial attributes and pose combination.

LSTM vs. Aligned LSTM: The performances of the baseline LSTM and the proposed aligned LSTM appear to be the same when only ρ is considered. However, MSE shows an improvement for the aligned LSTM, especially when the pose channel is included. This indicates that some improvement in performance could be achieved by using aligned channels. Both LSTMs perform the best for the facial attributes-pose combination yielding ρ value of 0.52 in both cases.

Multichannel attention LSTM: The proposed temporal attention-based LSTM network that selects only one (the most relevant) visual channel at every time step outperforms SVR in all cases, and other LSTMs in most cases in terms of both ρ and MSE. We experimented with using weighted versions of all three channels, instead of choosing the one with highest attention weight. This result was not superior to our current scheme of selecting only the best channel at a time step. Table 4 clearly shows that unlike LSTM or aligned LSTM, the performance of the attention network stands out when all three channels are added, for it achieves a correlation score of 0.57 - the best observed value in all possible scenarios. There is also a significant drop in the corresponding MSE values of attention-based LSTM, when all three channels are used. This supports our hypothesis that attention is an important contributor to how the visual attributes of a public speaker is perceived.

Visualizing attention: In addition to the quantitative results, we are interested in gaining further insights into the operation of the attention-based network. We thus look into the attention weights learned by the network at every time step to see if it corresponds to our intuitive understanding of attention to visual cues. Fig. 6 presents visualization re-

sults of the attention network for two sample videos from our database. The figure shows the sequence of the chosen visual channel (corresponding to the maximum attention weight) for each time step as generated by our network. This is computed at the volume-level (comprising 11 frames) as described earlier. In Fig. 6, block A shows which channel has the highest attention at every time step, while block B and C expand specific parts of the video to demonstrate the details of how attention works in our system. A representative frame from the corresponding part of the video is shown for easy validation. It is clear that when the speaker's face is shown on screen up-close, our attention network correctly selects the facial attributes channel as the most important cue. Likewise, when the full upper body is visible, attention switches to the human attributes channel, and when the speaker is at a distance, the pose channel is selected by the attention network. Also note that the frames with no speaker present in them are also correctly rejected by our system (white blocks). This visualization of the learned attention weights increases interpretability, and aligns with the intuitive human understanding of visual attention.

5.5. Discussion

From the various experimental results, we observe that human attributes and pose together form the weakest combination of visual cues. However, when the pose or the human attributes channel is combined with the facial attributes prediction accuracy improves. This could be intuitively explained by the fact that during TED talks, closeup of the speakers' faces appear on-screen more frequently, and perhaps influence the viewers' perception more than other cues. Hence combining facial attributes with any other visual cues yield better results. In other words, facial attributes are the most important among the visual cues we considered. We also notice that our attention-based LSTM performs the best when all three channels are included. This suggests that attention networks are particularly useful for dealing with larger number of channels, where it is natural to switch attention from one cue to another. We expect that our attention network will be further useful when the attention is sought among multiple modalities, say, visual, speech and lexical channels.

The visualization in Fig. 6 shows that the duration for which our system selects the human attributes channel (indicated by the red blocks) is less compared to the other two channels. This also aligns with our previous observation that the human attributes channel may be weaker than the others as these attributes do not change much over the duration of the video. This can also be seen in Table 4, where improvement in ρ between the face-pose and face-pose-human attributes is not very large, while MSE values remain the same.

A limitation of our framework lies in detecting the presence of a speaker in a video. As seen in Fig. 6, our system can correctly ignore the frames where no speaker is present. However, it also rejects the frames where more than one person detected, e.g., when faces are detected in the audience. Our current system can not distinguish between a speaker and any other human face shown in slides or in the audience (see Fig. 6, second example block B).

6. Conclusion

In this work, we have analyzed the visual behavior of speakers in the context of public speaking. We proposed a computational framework for predicting popularity of public speaking videos from visual cues related to physical appearance, facial expressions, and pose variations. Our framework learns to select the visual channel with maximum attention (learned as a latent variable) at every time step, and uses the sequence of the selected channels to predict video popularity. Note that the full architecture is trainable end-to-end; however, since the CNN streams require large amount of data to train, they are initialized with networks pretrained on appropriate databases with the respective sources being different. A large database comprising 1864 TED talk videos and their corresponding YouTube metadata was constructed to facilitate the study of public speaking in general. Extensive experiments on this database showed that our framework can predict video popularity from visual cues alone with significant accuracy. The proposed network learns a human-like selective attention mechanism depending on the visual cues present on screen. We observe that facial attributes contribute the most towards video popularity, while the human attributes channel gets much less attention in the network.

Our framework can be easily extended to include acoustic and lexical channels if prediction of video popularity is the primary objective. The prediction accuracy is also expected to improve significantly if verbal cues (e.g. speech prosody, intonation, fluency) are added to our framework. Due to the subjectivity involved in the very concept of video popularity, the popularity scores computed from YouTube ratings could be noisy. This is a common issue in most behavior modeling tasks requiring human annotations. However, how to quantify such measures better, especially for large scale data where seeking human annotation is expensive and time-consuming, is an open problem.

Acknowledgement

The authors would like to thank Research-I foundation, IIT Kanpur, and Nvidia Corp. for donating a Titan X GPU to support this research.

References

[1] S. Albert and J. M. Dabbs Jr. Physical distance and persuasion. *J of personality and social psychology*, 15(3):265, 1970.

[2] K. Cho, A. Courville, and Y. Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Trans Multimedia*, 17(11):1875–1886, 2015.

[3] J. Choi, Z. Wang, S. C. Lee, and W. J. Jeon. A spatio-temporal pyramid matching for video retrieval. *Computer Vision and Image Understanding*, 117(6):660–669, 2013.

[4] J. A. Daly, A. L. Vangelisti, and S. G. Lawrence. Self-focused attention and public speaking anxiety. *Personality and Individual Differences*, 10(8):903–913, 1989.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, 2007.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[7] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. R-cnns for pose estimation and action detection. 2014.

[8] S. Gonzalez and X. Anguera. Perceptually inspired features for speaker likability classification. In *ICASSP*, pages 8490–8494, 2013.

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] A. Kar, N. Rai, K. Sikka, and G. Sharma. Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[12] Z. Li, E. Gavves, M. Jain, and C. G. Snoek. Videolstm convolves, attends and flows for action recognition. *arXiv preprint arXiv:1607.01794*, 2016.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. *arXiv preprint arXiv:1512.02325*, 2015.

[14] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[15] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015.

[16] L. S. Nguyen, A. Marcos-Ramiro, M. Marrón Romera, and D. Gatica-Perez. Multimodal analysis of body communication cues in employment interviews. In *ICMI*, pages 437–444, 2013.

[17] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency. Deep multimodal fusion for persuasiveness prediction. In *Proc. ACM ICMI*, pages 284–288. ACM, 2016.

[18] D. P. Pertaub, M. Slater, and C. Barker. An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and virtual environments*, 11(1):68–78, 2002.

[19] R. E. Riggio and H. S. Friedman. Impression formation: The role of expressive behavior. *J Personality and Social Psychology*, 50(2):421, 1986.

[20] A. Rosenberg and J. Hirschberg. Acoustic/prosodic and lexical correlates of charismatic speech. In *Interspeech*, pages 513–516, 2005.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *Int J Computer Vision*, 115(3):211–252, 2015.

[22] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. In *CVPR*, pages 896–904, 2015.

[23] G. Sharma and F. Jurie. Learning discriminative spatial representation for image classification. In *BMVC*, pages 1–11, 2011.

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

[25] E. Strangert and J. Gustafson. What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations. In *Interspeech*, volume 8, pages 1688–1691, 2008.

[26] M. I. Tanveer, J. Liu, and M. E. Hoque. Unsupervised extraction of human-interpretable nonverbal behavioral cues in a public speaking scenario. In *ACM MM*, pages 863–866, 2015.

[27] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[28] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume 1, pages I–I. IEEE, 2001.

[29] C. J. Wesson. *The Communication and Influence of Confidence and Uncertainty*. PhD thesis, University of Wolverhampton, UK, 2005.

[30] T. Wörtwein, M. Chollet, B. Schauerte, L.-P. Morency, R. Stiefelhagen, and S. Scherer. Multimodal public speaking performance assessment. In *Proc. ACM ICMI*, pages 43–50, 2015.

[31] T. Wörtwein, L.-P. Morency, and S. Scherer. Automatic assessment and analysis of public speaking anxiety: A virtual audience case study. In *Proc. ACII*, pages 187–193, 2015.

[32] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[33] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.