2016 IEEE 16th International Conference on Data Mining

Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis

Soujanya Poria
Temasek Laboratories
Nanyang Technological University
Singapore
sporia@ntu.edu.sg

Iti Chaturvedi, Erik Cambria
School of Computer Science and Engineering
Nanyang Technological University
Singapore
{iti,cambria}@ntu.edu.sg

Amir Hussain
School of Natural Sciences
University of Stirling
United Kingdom
ahu@cs.stir.ac.uk

***Abstract*—Technology has enabled anyone with an Internet connection to easily create and share their ideas, opinions and content with millions of other people around the world. Much of the content being posted and consumed online is multimodal. With billions of phones, tablets and PCs shipping today with built-in cameras and a host of new video-equipped wearables like Google Glass on the horizon, the amount of video on the Internet will only continue to increase. It has become increasingly difficult for researchers to keep up with this deluge of multimodal content, let alone organize or make sense of it. Mining useful knowledge from video is a critical need that will grow exponentially, in pace with the global growth of content. This is particularly important in sentiment analysis, as both service and product reviews are gradually shifting from unimodal to multimodal. We present a novel method to extract features from visual and textual modalities using deep convolutional neural networks. By feeding such features to a multiple kernel learning classifier, we significantly outperform the state of the art of multimodal emotion recognition and sentiment analysis on different datasets.**

***Index Terms*—Multimodal sentiment analysis; Deep learning; Convolutional neural networks; Multiple kernel learning**

I. Introduction

Sentiment extraction from text has made considerable progress in the past few years [1], [2]. People, however, are gradually shifting from text to video to express their opinion about a product or service, as it is now much easier and faster for them to produce and share multimodal content [3]. For the same reasons, potential customers are now more inclined to browse for video reviews of the product they are interested in, rather than looking for lengthy written reviews [4]. Another reason for doing this is that, while trustable written reviews are quite hard to find, searching for good video reviews is as easy as typing the name of the product on YouTube and choosing the clips with more views [5].

This leads to the need for identifying sentiment and emotions from video as a source of multimodal information. However, there are major challenges which need to be overcome, e.g., expressiveness of opinion varies widely from person to person. Some people express their opinions more vocally, some more visually and others rely exclusively on logic and express little emotion. Furthermore, plenty of research has been conducted in the field of audio-visual emotion recognition.

Some recent work has also been conducted on fusing different modalities to detect emotions and polarity from videos [6], [7], [8]. This paper conducts extensive research on the different facets of this topic and aims to solve the following two questions:

1) Is a common framework useful for both multimodal emotion recognition and multimodal sentiment analysis?
2) Can audio, visual and textual features jointly enhance the performance of sentiment analysis classifiers?

In this paper, we propose a temporal convolutional neural network (CNN) where each pair of images at time t and $t+1$ are combined into a single image. Such a model is sensitive to sequence of images and learns a dictionary of features that are portable across languages. In a deep CNN, each hidden layer is obtained by convolving a matrix of weights with the matrix of activations at the layer below and the weights are trained using back propagation [9].

Furthermore, we have additional layers of recurrent neurons in the deep model. Recurrent neural networks (RNN) have feedback connections among neurons that can model dependencies in time sequences. Here, each hidden layer state is a function of the previous state, which can be further expanded as a function of all the previous states. In [10], the authors proposed convolutional RNNs to capture spatial structure information in static images. In contrast, our model uses RNN to capture spatial and temporal patterns that are inherent in video sequence. Our experiments showed that while using only RNN or deep CNN does not provide good classifications, combining the two models results in tremendous speed up and accuracy.

Multiple kernel learning (MKL) is a feature selection method where features are organized into groups and each group has its own kernel function [11]. MKL further improved our results, as it is able to combine data from different modalities effectively. Figure 1 illustrates the convolutional recurrent multiple kernel learning (CRMKL) model, which combines sentiment features in audio, video and text. In [12], the authors propose the use of a multi-resolution CNN to capture temporal features in YouTube videos. However, to our knowledge, this type of temporal CNN has not been previously used for sentiment analysis.

2374-8486/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDM.2016.178

IEEE computer society

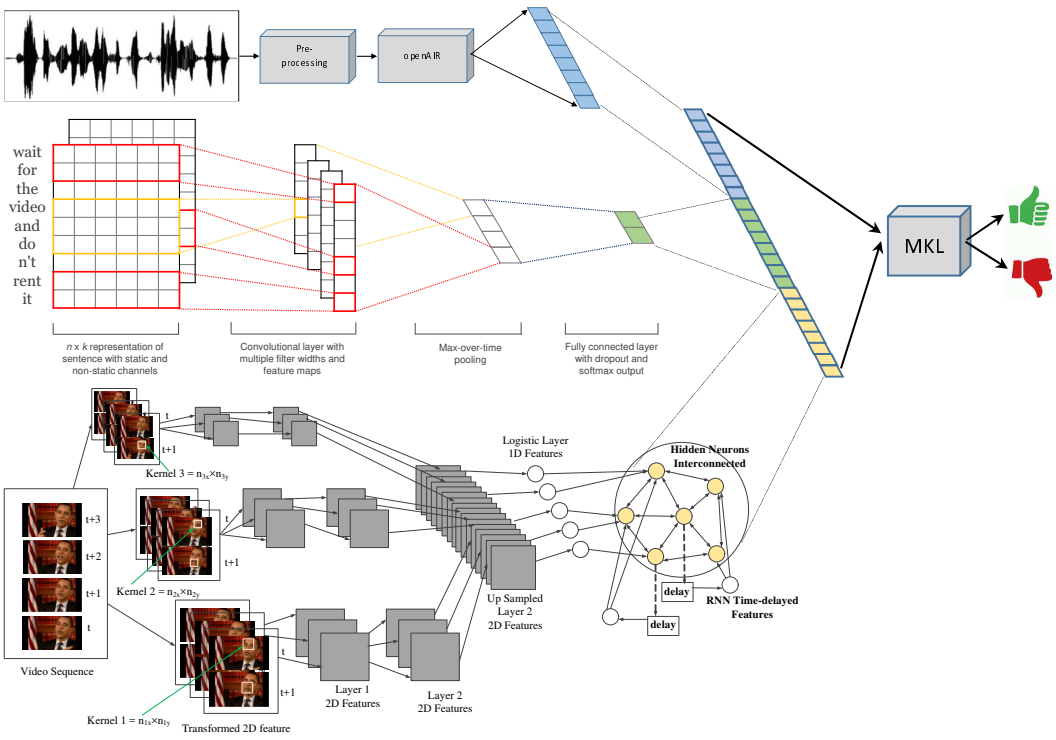

Fig. 1. The CRMKL model combining sentiment features in audio, video and text.

The organization of the paper is as follows: Section II reviews related works and datasets on multi-modal sentiment detection; Section III introduces the datasets used in our evaluation; Section IV provides the preliminary concepts necessary to understand the present work; Section V describes in detail the convolutional recurrent MKL framework for feature extraction and fusion from different modalities; finally, in Section VI we validate our method on different datasets.

II. Related Work and Contributions

Traditional methods could only classify text into different topics irrespective of user interests. In contrast, sentiment is an ordinal variable that can rank user interests in a sequential order. Sentiment prediction requires understanding of the sentence context and, hence, is a much more difficult task than topic classification. While sentiment prediction only identifies positive or negative customer experience, emotion recognition accounts for specific emotions, which can be used to create resonance among reviews for a certain product.

Recent work on text modality has used CNN for sentiment-related tasks such as sarcasm detection [13] and aspect-based opinion mining [14]. [15] developed a facial expression coding system (FACS) and six facial expressions that are able to provide sufficient clues to detect emotions. Recent studies on speech-based emotion analysis [4] have focused on identifying several acoustic features. One of the earliest works on fusing audio-visual emotion recognition [16] showed that a bimodal system yields higher accuracy than any unimodal system.

More recent research in audio-visual emotion recognition has been conducted at either feature level [17] or decision level [18]. Though there are plenty of research articles on audio-visual emotion recognition, only a few research works have been carried out on multimodal emotion recognition or sentiment analysis using textual clues with visual and audio modality. [5] and [19] fused information from audio, visual and textual modalities to extract emotion and sentiment.

[20] and [21] fused audio and textual modality at feature level for emotion recognition. [22] fused audio and textual clues at decision level. Similar to [23], this paper targets the classification of each sentence or utterance instead of the entire review. However, in [23] the authors generate user-defined feature-based summaries, which is not scalable in large datasets. Instead, we consider a deep neural network, where each layer automatically learns features in an unsupervised manner. This is followed by fine-tuning using a subset of known labels. In this way, the model is able to learn the lexicon of each new dataset during training. The following is a summary of the significance and contributions of this paper:

- The paper combines video, audio and text modality in order to effectively detect sentiment in a subject-independent manner. Our first contribution is that we have used MKL to fuse the three modalities. While the state of the art [24] uses a single kernel support vector machine (SVM) classifier to fuse all three modalities, we use multiple kernels to adapt to different modalities and, hence, achieve higher accuracy.

- Our second contribution is the novel integration of CNN with a low dimensional RNN, which is computationally much faster on large video data compared to baselines. In particular, for better modeling overlaps among features learned during temporal convolution, we consider distributed time-delayed features in the video. This can be achieved by initializing the weights of RNN with the covariance matrix of output feature vectors learned by the CNN.

III. Datasets Used

In this section, we describe the datasets used in the multimodal sentiment analysis and multimodal emotion recognition experiments. Our method can be easily used for the multi-class problem of neutral, positive, and negative sentiments. In this paper, we have followed previous authors on the benchmarks and excluded neutral reviews so that a simple comparison is possible. The method can also be used if one or two of the different modalities namely video, audio and text are present.

A. Multimodal Sentiment Analysis Dataset

We validated our method on three benchmark multimodal sentiment analysis datasets. The multimodal opinion utterances dataset (MOUD) was used to train the multimodal sentiment analysis module. The aim of the experiment was to predict the target label of each utterance in a video as positive or negative, where an utterance is a video segment of about 5 seconds.

For our experiment, we use the dataset developed by Morency et al. [25]. They collected videos from popular social media (e.g., YouTube) using several keywords (e.g., "favorite products") to produce search results consisting of videos of either product reviews or recommendation.

On average, each video has 6 utterances and each utterance is 5 seconds long. Each utterance in a video is annotated separately as positive or negative. Hence, sentiment can change during the course of a product review. The dataset contains 498 utterances labeled either positive, negative or neutral. In our experiment, we do not consider neutral labels, that leads to the final dataset consisting of 448 utterances.

Apart from the MOUD dataset, the trained model was then validated on YouTube and ICT-MMMO dataset. The former contains 110 negative and 87 positive videos of product reviews, the latter contains 230 positive and 119 negative videos. ICT-MMMO dataset, however, is not a utterance-level dataset. Hence, we manually split the videos into utterances.

B. Multimodal Emotion Recognition Dataset

The USC IEMOCAP database [26] was collected for studying multimodal expressive dyadic interactions. This dataset contains 12 hours of video data split into 5 minutes of dyadic interaction between professional male and female actors. Each interaction session was split into spoken utterances and labeled by at least 3 annotators into one emotion category, i.e., *happy, sad, angry, surprised, excited, frustration, disgust, fear* and other. The dataset contains 1,083 *angry*, 1,630 *happy*, 1,083 *sad*, and 1,683 neutral videos.

IV. Preliminaries

A. Deep Convolutional Neural Networks

A deep neural network can be viewed as a composite of simple, unsupervised models such as restricted Boltzmann machines (RBMs) where each hidden layer serves as the visible layer for the next RBM. RBM is a bipartite graph comprising two layers of neurons (a visible and a hidden layer), where the connections among neurons in the same layer are not allowed. Such a model can be first trained in an unsupervised manner, followed by fine-tuning using a subset of the data with known labels.

To train such a multi-layer system, we must compute the gradient of the total energy function E with respect to the weights in all the layers. To learn such weights and maximize the global energy function, the approximate maximum likelihood contrastive divergence approach can be used. This method employs each training sample to initialize the visible layer. Next, it uses the Gibbs sampling algorithm to update the hidden layer and then reconstruct the visible layer consecutively, until convergence. As an example, here we use a logistic regression model to learn the binary hidden neurons and each visible unit is assumed to be a sample from a normal distribution. The continuous state $\hat{h}_j$ of the hidden neuron j, with bias b_j, is a weighted sum over all continuous visible nodes $\boldsymbol{v}$ and is given by:

$$\hat{h}_j = b_j + \sum_i v_i w_{ij}, \tag{1}$$

where w_{ij} is the connection weight to hidden neuron j from visible node v_i. The binary state h_j of the hidden neuron can be defined by a sigmoid activation function:

$$h_j = \frac{1}{1 + e^{-\hat{h}_j}}, \tag{2}$$

Similarly, in the next iteration, the continuous state of each visible node v_i is reconstructed. Here, we determine the state of visible node i, with bias c_i, as a random sample from the normal distribution where the mean is a weighted sum over all binary hidden neurons and is given by:

$$v_i = c_i + \sum_j h_i w_{ij}, \tag{3}$$

where w_{ij} is the connection weight to hidden neuron j from visible node i. This continuous state is a random sample from $\mathcal{N}(v_i, \sigma)$, where σ is the variance of all visible nodes.

Unlike hidden neurons, visible nodes can take continuous values in a Gaussian RBM. Lastly, the weights are updated as the difference between the original and reconstructed visible layer labeled as the vector $\boldsymbol{v}_{recon}$ using:

$$\triangle w_{ij} = \alpha(< v_i h_j >_{data} - < v_i h_j >_{recon}), \tag{4}$$

where α is the learning rate and $< v_i h_j >$ is the expected frequency with which visible unit i and hidden unit j are active together when the visible vectors are sampled from the training set and the hidden units are determined by (1).

Finally, the energy of a deep neural network can be determined in the final layer using:

$$\mathrm{E} = -\sum_{i,j} v_i h_j w_{ij}, \tag{5}$$

To extend the deep neural network to a convolutional deep neural network, we simply partition the hidden layer into Z groups [27] . Each of the Z groups is associated with a $n_x \times n_y$ filter where n_x is the height of the kernel and n_y is the width of the kernel. Let us assume that the input image has dimension $\mathrm{L_x} \times \mathrm{L_y}$. Then, the convolution will result in a hidden layer of Z groups each of dimension $(\mathrm{L_x} - n_x + 1) \times (\mathrm{L_y} - n_y + 1)$. These learned kernel weights are shared among all hidden units in a particular group. The energy function of layer l is now a sum over the energy of individual blocks given by:

$$\mathrm{E}^l = -\sum_{z=1}^{\mathrm{Z}} \sum_{i,j}^{(\mathrm{L_x}-n_x+1),(\mathrm{L_y}-n_y+1)} \sum_{r,s}^{n_x,n_y} v_{i+r-1,j+s-1} h_{ij}^z w_{rs}^l. \tag{6}$$

Hence, each layer of a deep CNN is referred to as a convolution RBM (CRBM). In such a model, the lower layers learn abstract concepts and the higher layers learn complex features for subjective sentences.

B. Recurrent Neural Networks

The standard RNN output, $\mathbf{x}_l(t)$, at time step t for each layer l is calculated using the following equations :

$$\mathbf{x}_l(t) = f(W_{\mathrm{R}}^l.\mathbf{x}_l(t-1) + W_l.\mathbf{x}_{l-1}(t)) + W_{\mathrm{C}} \int_{t-k}^{t} \mathrm{x}_l(t)dt \tag{7}$$

where W_{R} is the interconnection matrix among hidden neurons and W_l is the weight matrix of connections between hidden neurons and the input nodes, $\mathbf{x}_{l-1}(t)$ is the input vector at time step t from layer $l-1$, vectors $\mathbf{x}_l(t)$ and $\mathbf{x}_l(t-1)$ represent hidden neuron activation at time steps t and $t-1$, respectively, and f is the non-linear activation function.

Furthermore, the distributed delays between output hidden features in each layer can be modeled via W_{C}. Unlike discrete time delays that can be learned separately for each hidden neuron, the distributed time delays are continuously changing due to the combined effect of different outputs and, hence, we use integration with respect to time to compute them.

In this paper, we propose to learn distributed time-delayed dependence using CNNs. Hence, a kernel of dimension $k \times k$ is able to capture distributed delays of up to k time points in the video sequence and can be approximated by the covariance matrix of features learned in the penultimate layer using (7). To learn the weights W_{R} of the RNN, back propagation through time is used where the hidden layer is unfolded in time using duplicate hidden neurons.

C. Multiple Kernel Learning

Consider a sequence of utterances $s(1), s(2), \ldots, s(T)$. The corresponding features for each utterance from audio, video and text data are denoted by $x(t)^a$, $x(t)^v$ and $x(t)^t$. MKL uses the corresponding target labels $y(t) \in \{+ve, -ve\}$ to optimize a dual form objective function with both min and max terms:

$$\max_{\beta} \min_{\alpha} \frac{1}{2} \sum_{i=1}^{T} \sum_{j=1}^{T} \alpha_i \alpha_j y(i) y(j) \Bigg(\sum_{m=1}^{M} \beta_m^a K_m^a(x(i)^a, x(j)^a) + \sum_{m=1}^{M} \beta_m^v K_m^v(x(i)^v, x(j)^v) + \sum_{m=1}^{M} \beta_m^t K_m^t(x(i)^t, x(j)^t) \Bigg) - \sum_{i=1}^{T} \alpha_i,$$

$$\text{s.t} \sum_{i=1}^{T} \alpha_i y(i) = 0, \ \sum_{m=1}^{\mathrm{M}} \beta_m = 1, \ 0 \le \alpha_i \le C \forall i. \tag{8}$$

where M is the total number of positive definite Gaussian kernels $K_m^a(x(i)^a, x(j)^a)$, $K_m^v(x(i)^v, x(j)^v)$ and $K_m^t(x(i)^t, x(j)^t)$ in each modality with a set of different parameters and α_i, b and $\beta_m \ge 0$ are coefficients to be learned simultaneously from the training data using quadratic programming.

V. CRMKL Model

In order to integrate RNN with CNN and MKL and to create the proposed CRMKL model, we extract features from audio, video and text and combine them using MKL. In particular, for video reviews we perform three steps: firstly, in order to capture temporal dependencies, we transform each pair of consecutive images at time t and $t+1$ into a single image; secondly, we include additional hidden layers of recurrent neurons in the deep CNN model; lastly, we initialize the distributed time-delay weight matrix of RNN with the covariance of CNN output.

A. Extracting Features from Visual Data

Sentiment analysis of large scale visual content can help to correctly extract sentiment of a topic. Deep CNNs have good accuracy on topic classification of videos, however they can get stuck in local minima on fine-grained problems such as sentiment and emotion detection [28], [29]. They are also extremely slow. Hence, we propose a layer of recurrent neurons to optimize the learning of features from video data.

Video sentiment detection faces two main challenges: firstly, it is an extremely computationally-expensive task; secondly, training datasets are weakly labeled and, hence, the trained model may not generalize well on new data. Since the video data is very large, we only consider every 10^{th} frame in our training videos. The constrained local model (CLM) is used to find the outline of the face in each frame [30]. The cropped frame size is further reduced by scaling down to a lower resolution. In this way, we can drastically reduce the amount of training video data.

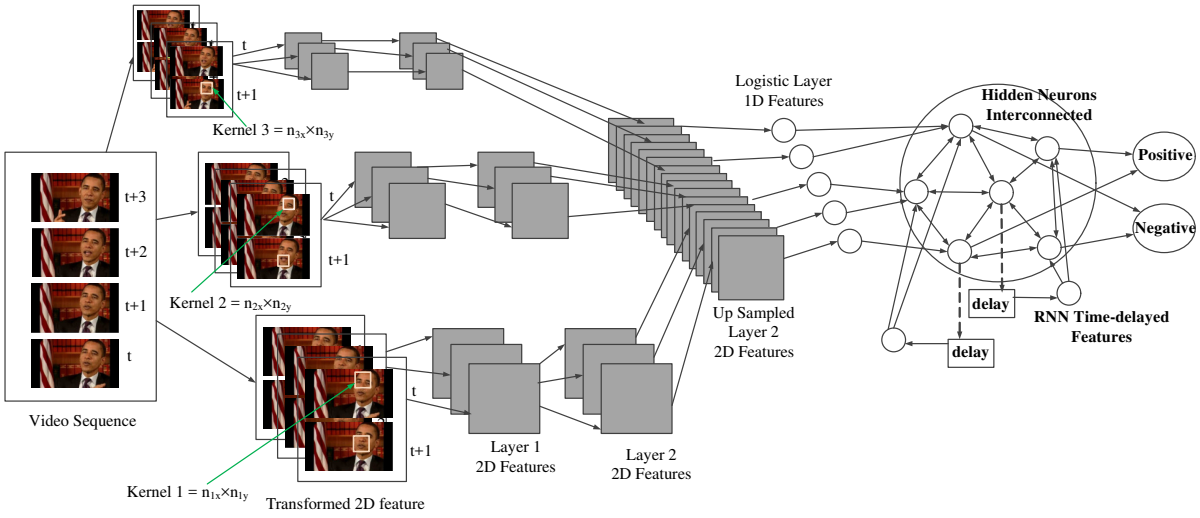

Fig. 2. Convolutional neural network for visual sentiment detection.

Figure 2 illustrates a convolutional RNN for visual sentiment detection. The input is a sequence of images in a video. To capture the temporal dependence, we transform each pair of consecutive images at t and $t+1$ into a single image. We use kernels of varying dimensions illustrated as Kernel 1, 2 and 3 to learn Layer 1 2D features from the transformed input.

Similarly, the second layer also uses kernels of varying dimensions to learn 2D features. Up sampling layer will transform features of different kernel sizes into uniform 2D features. Next, a logistic layer of neurons is used to prepare input for a RNN. Here, we have an inter-connected layer of neurons that can model long time delays using delay states. The final output layer classifies each video image as 'Positive' or 'Negative'.

In order to generalize the model to other domains, we train it using faces of different shapes and sizes. In order to validate it in a speaker-independent manner, moreover, we train the model on videos of product reviews in one domain and test on videos from a completely different domain. Pre-processing involved scaling all video frames to half the resolution. Each pair of consecutive video frames were converted into a single frame so as to achieve temporal convolution features. All the frames were standardized to 250 × 500 pixels by padding with zeros.

The first convolution layer contains 100 kernels of size 10×20, the next convolution layer had 100 kernels of size 20×30, this was followed by a logistic layer of 300 neurons and a recurrent layer of 50 neurons. The convolution layers were interleaved with pooling layers of dimension 2×2.

B. Extracting Features from Audio Data

We automatically extracted audio features from each annotated segment of the videos. Audio features were also extracted in 30Hz frame-rate and we used a sliding window of 100ms. To compute the features, we used the open source software openSMILE [31]. Specifically, this toolkit automatically extracts pitch and voice intensity. Z-standardization was used to perform voice normalization. Basically, voice normalization was performed and voice intensity was thresholded to identify samples with and without voice. The features extracted by openSMILE consist of several Low Level Descriptors (LLD) and statistical functionals of them. Some of the functionals are *amplitude mean, arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, interquartile ranges, linear regression slope*, etc. So, counting all functionals of each LLD, we obtained 6,373 features.

C. Extracting Features from Textual Data

We used a CNN as a trainable feature extractor to extract features from the textual data. Each utterance in the original dataset is in Spanish. While it is usually better to work directly with the source language, in this work we translated each utterance from Spanish to English using Google Translate. Without the translation into English, 68.56% accuracy was obtained on the MOUD dataset. The choice of CNN for feature extraction is justified by the following considerations: the CNN sentence model uses convolution as an operator to combine semantically-related word vectors and the convolution layers extract features in a hierarchical manner.

Each RBM layer is trained in an unsupervised manner and then the complete deep model can be fine-tuned using a subset of the dataset with known labels. The features learned in an unsupervised manner in each layer may not be the best for classification but can be used to train state-of-the-art classifiers such as SVM or Naïve Bayes.

Depending on the length of a sentence, the higher-order features can be short and focused or long and global spanning the entire sentence. CNNs form local features for each word and combine them to produce a global feature vector for the whole text using several hidden layers.

In this way, we can model semantic relations between words that may not be syntactically related in a parse tree. These features that the CNN builds internally can be extracted and used as input for another, more advanced classifier. In other words, this turns CNN, originally a supervised classifier, into a trainable feature extractor. To form the input for the CNN feature extractor, for each word in the text we constructed a 306-dimensional vector by concatenating two parts:

- *Word embeddings*: We used a publicly available word2vec dictionary [32], which has been trained on a 100-million word corpus from Google News using the continuous bag-of-words architecture. This dictionary provides a 300-dimensional vector for each word. For words not found in this dictionary, we used random vectors.
- *Part of speech*: We used 6 basic parts of speech (noun, verb, adjective, adverb, preposition, conjunction) encoded as a 6-dimensional binary vector. We used Stanford Tagger as a part of speech tagger [33].

If an instance s has n words then we represent the input vector for that instance $s_{1:n} = s_1 \oplus s_2 \oplus ... \oplus s_n$. Here, $s_i \in \Re^k$ is a k dimensional feature vector for word s_i (in this case k=306). In our experiments, all texts were very short, consisting of one sentence, the longest one being of 65 words. Thus, all input vectors were of dimension 306 × (2 + 65 + 2) = 21,114. The CNN we used consisted of 7 layers:

- *Input layer* of 21,114 neurons.
- *Convolution layer* with a kernel size of 3,4 and 50 feature maps each. The output of the layer was computed with a non-linear function; we used the ReLU.
- *Max-pool layer* with max-pool size of 2. Max pooling operation over the feature map will take the maximum value as the feature corresponding to a particular kernel vector and, hence, discard highly similar features during convolution.
- *Convolution layer* of kernel size of 2, 100 feature maps, also using the ReLU.
- *Max-pool layer* with max-pool size of 2.
- *Fully-connected layer* of 500 neurons. The values of these neurons were later used as the extracted features. For regularization, we employ dropout on the penultimate layer with a constraint on L2-norms of the weight vectors.
- *Output softmax layer* of 2 neurons. The final layer which outputs two labels: positive or negative.

The features were extracted from the penultimate fully-connected layer of the CNN. In this way, we used the last output layer of the CNN only for training, but for actual decision-making, we replaced it with more sophisticated classifiers such as SVM or MKL.

On MOUD dataset, using only CNN as a classifier, 75.50% was obtained which is in fact lower than the result (79.77%) obtained when CNN was used to extract trainable features for the SVM classifier (Table I). We also tried other word vectors having different dimensions, e.g., Glove word vectors and Collobart's word vectors.

However, the best accuracy was obtained using Google word2vec. We would like to clarify that we only translate the text form of utterances into English. The audio and video data is however in Spanish. The purpose of translation is to leverage on the lexical resources in English and to interpret the emotions in the video and audio with text.

D. Feature Selection and Fusion

We significantly reduced the number of features using feature selection. We used two different feature selectors: one based on the cyclic correlation-based feature subset selection (CFS) and another based on principal component analysis (PCA). The main idea of CFS is that useful feature subsets should contain features that are highly correlated with the target class while being uncorrelated with each other. PCA is a slightly different method, that uses an orthogonal transformation to convert the data into a set of variables that are linearly uncorrelated called principal components. The components can be ranked by their magnitude in the data. By discarding smaller (less meaningful) components, PCA allows for dimensionality reduction and analogical reasoning [34]. Here, we select top K features from each method, where K was experimentally determined by trial. For example, in case of audio, visual and textual fusion, K was set to 300.

Feature selection for multimodal sentiment and emotion analysis is done using MOUD and IEMOCAP training dataset, respectively. However, for each unimodal, each bimodal, and the multimodal experiment, feature selection is done separately. Feature-level fusion is achieved by concatenation of the feature vectors obtained for each of the three modalities.

Clearly, the combined feature vectors from different modalities are heterogeneous in nature. Hence, the resulting vectors, along with the corresponding sentiment polarity labels from the training set, were used to train a classifier with a MKL algorithm; we used the SPF-GMKL implementation [35], which is designed to deal with heterogeneous data.

The parameters of the classifier were found by cross-validation. We chose a configuration with 8 kernels: 5 RBF with gamma from 0.01 to 0.05 and 3 polynomial with powers 2, 3, 4. We also tried Simple-MKL; it gave slightly lower results.

E. Computational Complexity

The computational complexity for a convolutional layer l is given by $O(n_{l-1}.s_l^2.n_l.m_l^2)$, where n_{l-1} and n_l are the number of input and output feature maps, $s_l = n_x^{l-1} \times n_y^{l-1}$ and $m_l = n_x^l \times n_y^l$ are the dimensions of the input and output feature maps. The computational complexity of a layer of recurrent hidden neurons is only $O(\mathrm{R} \times n^2)$, where R is the maximum time delay considered and n is the number of neurons.

We can hence conclude that the computational complexity of RNN is much lower than CNN for each iteration of training. Therefore, in this paper we first train CNN for a limited number of epochs and then the partially learned features are further evolved using a low dimensional RNN for video data.

TABLE I
ACCURACY OF STATE-OF-THE-ART METHOD COMPARED WITH OUR METHOD WITH FEATURE-LEVEL FUSION ON MOUD DATASET. THE NUMBER OF FEATURES REFERS TO OUR EXPERIMENTS, NOT TO [24].

		Text	Visual	Audio	[24]	Our method	
						without feature selection	with feature selection
# features, without selection		500	50	6373		↰	
Unimodal	# features, with selection	437	–	–	70.94%	79.14%	79.77%
		–	50	–	67.31%	94.50%	94.50%
		–	–	325	64.85%	74.49%	74.22%
Bimodal		381	50	–	72.39%	95.75%	96.21%
		384	–	81	72.88%	83.85%	84.12%
		–	50	217	68.86%	95.38%	95.68%
Multimodal		50	89	64	74.09%	96.12%	96.55%

VI. EXPERIMENTAL RESULTS AND DISCUSSION

We used a common framework for both sentiment and emotion detection. For multimodal sentiment analysis, following Perez et al., we have used the entire set of 448 utterances in MOUD dataset and run ten-fold cross-validation using CRMKL. In addition, to test the generalization ability of the model on new datasets, we have also shown results on test data from YouTube and ICT-MMMO. For comparison with unimodal datasets such as only video or only text, we have used SVM as a baseline classifier. For the case of emotion recognition, that is a much more fine-grained problem than sentiment detection, we evaluate our model via ten-fold cross-validation on IEMOCAP dataset. Feature selection was not done for visual modality as the deep CNN module in CRMKL automatically learns the best features. Our experiments showed that feature selection on visual data can lead to reduction in accuracy.

Table I shows the 10-fold cross-validation results obtained on MOUD dataset. The visual module of CRMKL, obtained 27% higher accuracy than the state of the art. When all modalities were used, 96.55% accuracy was obtained outperforming the state of the art by more than 20%. Next, to assess the accuracy of the model on an unknown dataset, we trained the model on MOUD dataset and tested on ICT-MMMO and YouTube dataset. On both of these datasets, the model performed notably well.

The visual classifier trained on the MOUD obtained 93.60% accuracy. Other unimodal classifiers did not perform well like the visual classifier in the cross-domain analysis. As ICT-MMMO dataset is a video-level sentiment dataset, utterance-level sentiment evaluation is not possible. Hence, after the model generates sentiment labels of all utterances for a video, we took the majority sentiment label of these utterances in order to label the video by its sentiment.

We got 85.30% accuracy on the ICT-MMMO dataset using the trained visual sentiment model on the MOUD dataset. The obtained accuracy on the ICT-MMMO dataset was lower than the other two datasets. This is because ICT-MMMO dataset was manually segmented into utterances and, hence, it is likely to have more noise compared to other datasets.

Not only the visual features, textual features are also novel as they indeed boosted the accuracy of the experiments where textual modality was involved. The unimodal experiment with only textual features outperformed the performance of the state of the art as shown in Table I. On all three datasets, the visual and textual modalities when combine together produced better accuracy than other bimodal experiments.

For multimodal emotion analysis, we used the same framework as we employed for multimodal sentiment analysis. The accuracy for all unimodal, bimodal and trimodal experiments are significantly better than the state of the art. However, the performance is not as good as the multimodal sentiment analysis experiments. One of the possible reasons for this is the use of same CNN configurations for both visual and textual sentiment feature extraction.

This raises the question of using larger number of neurons and layers in CRMKL for visual and textual emotion feature extraction. This is of course a fundamental task of our future work. The following observations were made from the multimodal emotion analysis experiments:

- We realized that the textual classifier recognized *angry*, *happy* and neutral instances well. However, *angry* and *sad* instances are very tough to distinguish from each other using textual clues. One of the possible reasons is that both classes are negative and many similar words are used to express them.
- In the case of audio modality, we observed better accuracy than textual modality for *sad* and neutral classes but not for *happy* and *angry* classes. The classifier misclassified many *happy* instances into *angry*. However, the classifier performed very well to discriminate between sadness and anger. We also observed that some *happy* instances were classified as neutral.
- Visual modality produced the best accuracy compared to the other two modalities. Though *angry* and *sad* faces can be effectively classified, the classifier showed some confusion between *angry* and *sad* faces. Neutral classes were also separated more accurately in respect to other classes though high confusion was observed between *happy* and *neutral* faces.

TABLE II
ACCURACY ON TEXTUAL (T), VISUAL (V), AUDIO (A) MODALITY AND COMPARISON WITH THE STATE OF THE ART.

Modalities		Emotion, on IEMOCAP			
		angry	*happy*	*sad*	neutral
T	**our results**	60.01%	58.71%	57.15%	61.25%
	state of the art	63.10%[1]	49.60%[1]	42.00%[1]	39.50%[1]
V	**our results**	69.50%	67.34%	67.41%	71.55%
	state of the art	41.80%[1]	63.60%[1]	52.60%[1]	47.00%[1]
A	**our results**	59.83%	56.81%	60.75%	67.91%
	state of the art	66.10%[1]	53.90%[1]	65.50%[1]	58.10%[1]
T + V	**our results**	74.81%	69.22%	74.85%	77.49%
	state of the art	–	–	–	–
T + A	**our results**	62.50%	65.21%	63.30%	69.25%
	state of the art	77.80%[1]	63.20%[1]	68.30%[1]	60.40%[1]
V + A	**our results**	71.86%	69.35%	74.23%	77.58%
	state of the art	–	–	–	–
A + V + T	**our results**	**79.20%**	**72.22%**	**75.63%**	**80.35%**
	state of the art	78.10%[1]	69.20%[1]	67.10%[1]	63.00%[1]

[1]by [36]

When we fused the modalities using feature-level fusion strategy, higher accuracy was obtained as compared to uni-modal classifiers, as expected. Although the identification accuracy has been improved for every emotion, the confusion between *sad* and *angry* face was still high.

The comparison with the state of the art (Table II) in terms of accuracy shows that the proposed method performed significantly better. For *sad* and neutral emotion the proposed method outperformed the state of the art by a margin of 8% and 17%, respectively. However, for *angry* and *happy* the performance is just slightly better.

Paired t-test showed statistical significance of all experiments with confidence level 95%. It can be found from the Table II that visual and textual modalities performed notably better than the state of the art. With the help of these two modalities, the proposed method outperformed the state of the art.

In this paper, we proposed the novel integration of CNN with a low dimensional RNN that can converge to the global maxima much faster than baselines. Hence, in Table II, for the emotion *sad*, the performance of visual modality and the bimodal combinations of visual modality with text and audio, respectively, is over 70%.

This is due to the superior performance of the proposed video classifier. In contrast, the bimodal combination of audio and text has a 10% lower accuracy that is similar to the baseline. The combined multimodal classifier of audio, video and text is slightly better than visual modality. This is because the video classifier dominates over the other two modalities.

Deep CNN have recently shown good performance on audio, video and text classification. Instead of using a single large hidden layer of neurons, deep models have several small layers of hidden neurons. Since each layer is independent, this results in tremendous reduction in complexity. Therefore, in this paper we construct a deep CNN for each modality, namely: audio, video, and text.

The groups of features learned by each of the three deep CNN are combined using MKL. In this way, we can reduce the number of input dimensions and group the features for MKL.

A. *Effect of Number of Hidden Layers*

Deep learning is able to approximate very long time-delays in video data via a hierarchy of hidden layers, where the features learned in one layer become the input data to the next layer. To determine the number of hidden layers of recurrent neurons, we consider the root mean square error (MSE) on training data. MSE is the cost function that the deep model is trying to minimize while learning the weights. Hence, this is a suitable metric for the improvement made by each hidden layer in a deep model.

Figure 3 reports the decrease in MSE with increasing number of hidden layers for the YouTube test dataset. It was also observed that the variance over 10-fold cross-validation reduces with increasing number of hidden layers. Hence, we can conclude that deep learning is suitable for extracting sentiments and emotions from video data. Since each layer is learned independent of the previous layer, the number of parameters is small and overfitting is avoided.

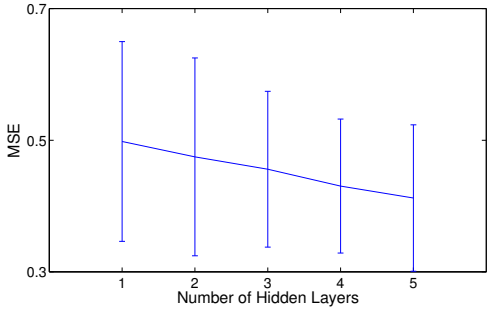

Fig. 3. MSE with respect to number of hidden layers.

B. Tuning of Hyper-parameters

As a performance measure, we adopted the F-score. Each dataset is split into training set, validation set, and test set. For all three modalities and for each hidden layer we consider different number of hidden units (i.e., n=50/200/500/700) and 5000 epochs of CNN training using the Theano based stochastic gradient descent. The number of hidden neurons in each layer is gradually increased until performance saturates due to overfitting. In particular, early overfitting occurs for the MOUD dataset.

Our best results are obtained with an ensemble of CNNs by 10-fold cross-validation that differ in their random initialization and mini-batches of 100 samples. Results on CNNs of various depths and sizes shows that deep CNN outperforms single-layer CNN with approximately the same number of parameters, which quantitatively validates the benefits of deep networks over shallow ones.

We see a consistent improvement as we use deeper models. Following previous authors, the word vector length was empirically set to 300, and unknown words were randomly initialized to vectors from Gaussian distributions. The 6 dimensional vector corresponds to 6 different parts of speech such as noun and verb.

C. Visualization of Features

The deep temporal CNN model automatically learns features from the training data, so that each neuron learns a specific feature such as eyes or mouth. In the first layer, the features learned are parts of the face and their sentiments, and the higher layers will combine these emotional features to learn the complete face and corresponding positive or negative sentiment. We visualize the feature detectors in the first layer of the network trained on the MOUD sentiment data.

We rank all image segments in the training data according to the activation of each detector. Figure 4 shows the top image segments activated at two feature detectors in the first layer of a deep CNN. We find that similar features such as eyes or mouth are expressed at the same hidden neuron. The feature detectors learn to recognize not only the part of the face but also the sentiment associated with it.

VII. Conclusion

Communication across the World Wide Web is rapidly shifting from unimodal data, i.e., text, to multimodal data, i.e., video. Extracting emotions and polarity from videos is hence becoming increasingly important for tasks such as social media marketing, brand positioning, and financial prediction.

In this paper, we proposed the fusion of speech, voice tone, and facial expressions for multimodal emotion recognition and sentiment analysis. In particular, we described a novel temporal deep convolutional neural network for visual and textual feature extraction and used multiple kernel learning to fuse heterogeneous features extracted from different modalities, namely: audio, video, and text.

In the future, we will focus on improving the accuracy of emotion detection via different neural network configurations. We will also consider annotation of ICT-MMMO dataset at utterance level for smoother training of the model.

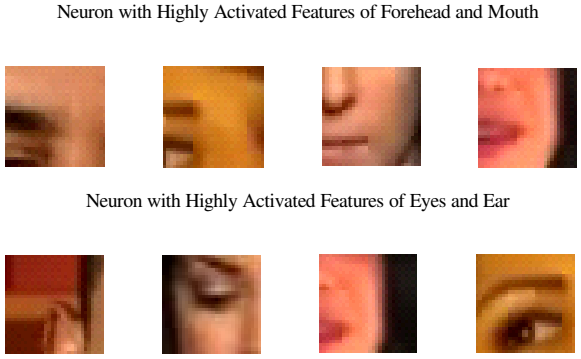

Fig. 4. Top image segments activated at two feature detectors in the first layer of deep CNN

References

[1] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 48–57, 2014.

[2] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[3] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1805–1812.

[4] D. Datcu and L. Rothkrantz, "Semantic audio-visual data fusion for automatic emotion recognition," *Euromedia'2008*, 2008.

[5] V. Rosas, R. Mihalcea, and L.-P. Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 0038–45, 2013.

[6] S. Poria, E. Cambria, A. Hussain, and G.-B. Huang, "Towards an intelligent framework for multimodal affective data analysis," *Neural Networks*, vol. 63, pp. 104–116, 2015.

[7] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, 2016.

[8] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[10] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, "Convolutional recurrent neural networks: Learning spatial dependencies for image representation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 18–26.

[11] N. Subrahmanya and Y. Shin, "Sparse multiple kernel learning for signal processing applications," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 5, pp. 788–798, 2010.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[13] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "A deeper look into sarcastic tweets using deep convolutional neural networks," in *COLING*, 2016.

[14] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowledge-Based Systems*, vol. 108, 2016.

[15] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Stanford University, Palo Alto, 1977.

[16] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Information, Communications and Signal Processing, 1997. ICICS., Proceedings of 1997 International Conference on*, vol. 1. IEEE, 1997, pp. 397–401.

[17] D. Datcu and L. J. Rothkrantz, "Emotion recognition using bimodal data fusion," in *Proceedings of the 12th International Conference on Computer Systems and Technologies*. ACM, 2011, pp. 122–128.

[18] B. Schuller, "Recognizing affect from linguistic information in 3d continuous space," *Affective Computing, IEEE Transactions on*, vol. 2, no. 4, pp. 192–205, 2011.

[19] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Speech language & multimedia technol., raytheon bbn technol., cambridge, ma, usa," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–4.

[20] A. Metallinou, S. Lee, and S. Narayanan, "Audio-visual emotion recognition using gaussian mixture models for face and voice," in *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*. IEEE, 2008, pp. 250–257.

[21] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 7–19, 2010.

[22] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *Affective Computing, IEEE Transactions on*, vol. 2, no. 1, pp. 10–21, 2011.

[23] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[24] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proceedings of ACL 2013*, 2013, pp. 973–982.

[25] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 2011, pp. 169–176.

[26] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[27] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09, 2009, pp. 609–616.

[28] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *EMNLP*, 2015, pp. 2539–2544.

[29] I. Chaturvedi, Y.-S. Ong, I. Tsang, R. Welsch, and E. Cambria, "Learning word dependencies in text by means of a deep recurrent belief network," *Knowledge-Based Systems*, vol. 108, pp. 144–154, 2016.

[30] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recogn.*, vol. 41, no. 10, pp. 3054–3067, 2008.

[31] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[32] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[33] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *IN PROCEEDINGS OF HLT-NAACL*, 2003, pp. 252–259.

[34] L. Oneto, F. Bisio, E. Cambria, and D. Anguita, "Statistical learning theory and ELM for big social data analysis," *IEEE Computational Intelligence Magazine*, vol. 11, no. 3, pp. 45–55, 2016.

[35] A. Jain, S. Vishwanathan, and M. Varma, "Spf-gmkl: generalized multiple kernel learning with a million kernels," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 750–758.

[36] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of svm trees for multimodal emotion recognition," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–4.