# GeThR-Net: A Generalized Temporally Hybrid Recurrent Neural Network for Multimodal Information Fusion

Ankit Gandhi[1(✉)], Arjun Sharma[1], Arijit Biswas[2], and Om Deshmukh[1]

[1] Xerox Research Centre India, Bengaluru, India
ankit.g1290@gmail.com, arjunsharma.iitg@gmail.com, om.deshmukh@xerox.com
[2] Amazon Development Center India, Chennai, India
arijitbiswas87@gmail.com

**Abstract.** Data generated from real world events are usually temporal and contain multimodal information such as audio, visual, depth, sensor etc. which are required to be intelligently combined for classification tasks. In this paper, we propose a novel generalized deep neural network architecture where temporal streams from multiple modalities are combined. There are total M+1 (M is the number of modalities) components in the proposed network. The first component is a novel temporally hybrid Recurrent Neural Network (RNN) that exploits the complimentary nature of the multimodal temporal information by allowing the network to learn both modality specific temporal dynamics as well as the dynamics in a multimodal feature space. M additional components are added to the network which extract discriminative but non-temporal cues from each modality. Finally, the predictions from all of these components are linearly combined using a set of automatically learned weights. We perform exhaustive experiments on three different datasets spanning four modalities. The proposed network is relatively 3.5 %, 5.7 % and 2 % better than the best performing temporal multimodal baseline for UCF-101, CCV and Multimodal Gesture datasets respectively.

## 1 Introduction

Humans typically perceive the world through multimodal sensory information [30] such as visual, audio, depth, etc. For example, when a person is running, we recognize the event by looking at how the body posture of the person is changing with time as well by listening to the periodic sound of his/her footsteps. Human brains can seamlessly process multimodal signals and accurately classify an event or an action. However, it is a challenging task for machines to exploit the complimentary nature and optimally combine multimodal information.

Recently, deep neural networks have been extensively used in computer vision, natural language processing and speech processing. LSTM [9], a Recurrent Neural Network (RNN) [35] architecture, has been extremely successful in

A. Gandhi and A. Sharma—Equally contributed.

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part II, LNCS 9914, pp. 883–899, 2016.
DOI: 10.1007/978-3-319-48881-3_58

temporal modelling and classification tasks such as handwriting recognition [8], action recognition [2], image and video captioning [4,31,44] and speech recognition [6,7]. RNNs can also be used to model multimodal information. These methods fall under two broad categories: (a) Early-Fusion: modality specific features are combined to create a feature representation and fed into a LSTM network for classification. (b) Late-Fusion: each modality is modelled using individual LSTM networks and their predictions are combined for classification [40]. Since early-fusion techniques do not learn any modality specific temporal dynamics, they fail to capture the discriminative temporal cues present in each modality. On the other hand, late-fusion methods cannot extract the discriminative temporal cues which might be available in a multimodal feature representation. In this paper, we propose a novel generalized temporally hybrid Recurrent Neural Network architecture called GeThR-Net which models the temporal dynamics of individual modalities (late fusion) as well as the overall temporal dynamics in a multimodal feature space (early fusion).

GeThR-Net has one temporal and $M$ ($M$ is the total number of modalities) non-temporal components. The novel temporal component of GeThR-Net models the long-term temporal information in a multimodal signal whereas the non-temporal components take care of situations where explicit temporal modelling is difficult. The temporal component consists of three layers. The first layer models each modality using individual modality-specific LSTM networks. The second layer combines the hidden representations from these LSTMs to form a multimodal feature representations corresponding to each time step. In the final layer, one multimodal LSTM is trained on the multimodal features obtained from the second layer. The output from the final layer is fed into a softmax layer for category-wise confidence prediction. We observe that in many real world scenarios, the temporal modelling of individual or multimodal information is extremely hard due to the presence of noise or high intra-class temporal variation. We address this issue by introducing additional $M$ components to GeThR-Net which model modality specific non-temporal cues by ignoring the temporal relationship across features extracted from different time-instants. The predictions corresponding to all $M + 1$ components in the proposed network are combined using a weighted vector learned from the validation dataset. We note that GeThR-Net can be used with any kind of modality information without any restriction on the number of modalities.

The main contributions of this paper are:

- We propose a generalized deep neural network architecture called GeThR-Net that could intelligently combine multimodal temporal information from any kind and from any number of streams.
- Our objective is to propose a general framework that could work with modalities of any kind. We demonstrate the effectiveness and wide applicability of GeThR-Net by evaluation of classification performance on three different action and gesture classification tasks, UCF-101 [28], Multimodal Gesture [5] and Columbia Consumer videos [13]. Four different modalities such as audio, appearance, short-term motion and skeleton are considered in our experiments.

We find out that GeThR-Net is relatively 3.5 %, 5.7 % and 2 % better than the best temporal multimodal baseline for UCF-101, CCV and Multimodal Gesture datasets respectively.

The full pipeline of the proposed approach is shown in Fig. 1. We discuss the relevant prior work in Sect. 2 followed by the details of GeThR-Net in Sect. 3. The details of experimental results are provided in Sect. 4.

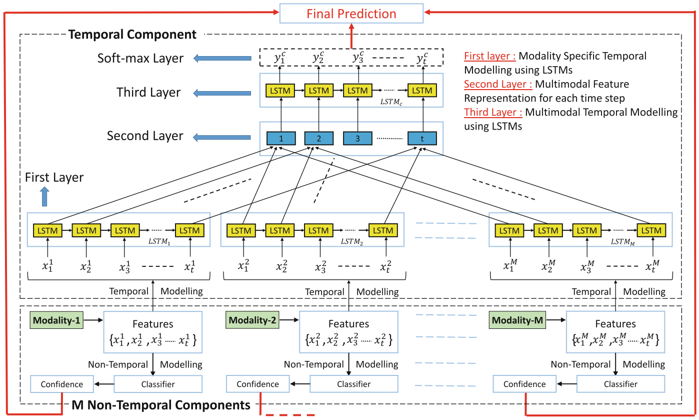
Final Prediction
Temporal Component
Soft-max Layer
$y_1^c$ $y_2^c$ $y_3^c$ ----- $y_t^c$
Third Layer
LSTM LSTM LSTM ...... LSTM
$LSTM_c$
Second Layer
1 2 3 ........... t
First Layer
First layer : Modality Specific Temporal Modelling using LSTMs
Second Layer : Multimodal Feature Representation for each time step
Third Layer : Multimodal Temporal Modelling using LSTMs
LSTM LSTM LSTM ...... LSTM
$LSTM_1$
$x_1^1$ $x_2^1$ $x_3^1$ ----- $x_t^1$
$LSTM_2$
$x_1^2$ $x_2^2$ $x_3^2$ ----- $x_t^2$
$LSTM_M$
$x_1^M$ $x_2^M$ $x_3^M$ ----- $x_t^M$
Temporal Modelling
Modality-1
Features $\{x_1^1, x_2^1, x_3^1 ..... x_t^1\}$
Modality-2
Features $\{x_1^2, x_2^2, x_3^2 ..... x_t^2\}$
Modality-M
Features $\{x_1^M, x_2^M, x_3^M ..... x_t^M\}$
Non-Temporal Modelling
Confidence
Classifier
M Non-Temporal Components

**Fig. 1.** The overall pipeline of the proposed approach GeThR-Net. The input to the system is a multimodal stream (e.g.: appearance, short-term motion, skeleton and/or audio for action/gesture classification tasks) and output is the class label. The proposed network has total $M + 1$ components ($M$ is the total number of modalities). The first component is a temporally hybrid network that models the modality specific temporal dynamics as well as the temporal dynamics in a multimodal feature space. Corresponding to each of the $M$ modalities, there is also a non-temporal classification component in the network. All of these components in the network are trained in an end-to-end fashion.

## 2 Related Work

In this section, we describe the relevant prior work on generic multimodal fusion and multimodal fusion using deep learning.

**Multimodal Information Fusion:** A good survey of different fusion strategies for multimodal information is in [1]. We discuss a few relevant papers here. The authors in [41] provide a general theoretical analysis for multimodal information fusion and implements novel information theoretic tools for multimedia applications. [37] proposes a two-step approach for an optimal multimodal fusion, where in the first step statistically independent modalities are found from raw features and in the second step, super-kernel fusion is used to find the optimal combination of individual modalities. In [10], the authors propose a method for detecting complex events in videos by using a new representation, called bi-modal words, to explore the representative joint audio and visual patterns. [12] proposes a method to extract a novel representation, the Short-term Audio-Visual Atom (S-AVA), for improved semantic concept detection in videos. The authors in [45] propose a rank minimization method to fuse the predicted confidence scores of multiple models based on different kinds of features. Their goal is to find a shared rank-2 pairwise relationship matrix (for the test samples) based on which each original score matrix from individual model can be decomposed into the common rank-2 matrix and sparse deviation errors. [26] proposes an early and a late fusion scheme for audio, visual and textual information fusion for semantic video analysis and demonstrates that the late fusion method works slightly better. In [22], the authors propose a multimodal fusion technique and describe a way to implement a generic framework for multimodal emotion recognition.

**Deep Learning for Multimodal Fusion:** In [20], the authors propose a deep autoencoder network that is pretrained using sparse Restricted Boltzmann Machines (RBM). The proposed method is used to learn multimodal feature representation for the task of audio-visual speech recognition. The authors in [29], propose a Deep Boltzmann Machine (DBM) for learning a generative model of data that consists of multiple and diverse input modalities. [27], proposes a multimodal representation learning framework that minimizes the variation information between data modalities through shared latent representations. In [38], the authors propose a unified deep neural network, which jointly learns feature relationships and class relationships, and simultaneously carries out video classification within the same framework utilizing the learned relationships. [17,18] proposes an approach for generating novel image captions given an image. This approach directly models the probability distribution of a word given previous words and an image using a network that consists of a deep RNN for sentences and a deep CNN for images. [36] proposes a novel bi-modal dynamic network for gesture recognition. High level audio and skeletal joints representations, extracted using dynamic Deep Belief Networks (DBN), are combined using a layer of perceptron. However, none of these approaches use RNNs for both multimodal and temporal data fusion and hence cannot learn features which truly represent the complimentary nature of multimodal features along the temporal dimension. The authors in [3], propose a multi-layer RNN for multi-modal emotion recognition. However, the number of layers in the proposed architecture is equal to the number of modalities, which restricts the maximum number of modalities which can be used simultaneously. The authors in [40] propose a hybrid deep learning framework for video classification that can model static spatial information, short-term motion, as well as long-term temporal clues in the videos. The spatial and the short-term motion features extracted from CNNs are combined using a regularized feature fusion network. LSTM is used to model only the modality specific long-term temporal information. However, in the proposed GeThR-Net, the temporally hybrid architecture can automatically combine temporal information from multiple modalities without requiring any explicit feature fusion framework. We also point out that unlike [40], in GeThR-Net, the multimodal fusion is performed at the LSTM network level.

To the best of authors' knowledge, there are no prior approaches where multimodal information fusion is performed at the RNN/LSTM level. GeThR-Net is the first method to use a temporally hybrid RNN which is capable of learning features from modalities of any kind without any upper-bound on the number of modalities.

## 3 Proposed Approach

In this section, we provide the details of the proposed deep neural network architecture GeThR-Net. First, we discuss how LSTM networks usually work. Next, we provide the descriptions of the temporal and non-temporal components of our network followed by how we combine predictions from all these components.

### 3.1 Long Short Term Memory Networks

Recently, a type of RNN, called Long Short Term Memory (LSTM) Networks, have been successfully employed to capture long-term temporal patterns and dependencies in videos for tasks such as video description generation, activity recognition etc. RNNs [35] are a special class of artificial neural networks, where cyclic connections are also allowed. These connections allow the networks to maintain a memory of the previous inputs, making them suitable for modelling sequential data. In LSTMs, this memory is maintained with the help of three non-linear multiplicative gates which control the in-flow, out-flow, and accumulation of information over time. We provide a detailed description of RNNs and LSTM networks below.

Given an input sequence $\mathbf{x} = \{x_t\}$ of length $T$, the fixed length hidden state or memory of an RNN $\mathbf{h}$ is given by

$$h_t = g(x_t, h_{t-1}) \quad t = 1, \ldots, T \tag{1}$$

We use $h_0 = 0$ in this work. Multiple such hidden layers can be stacked on top of each other, with $x_t$ in Eq. 1 replaced with the activation at time $t$ of the previous hidden layer, to obtain a 'deep' recurrent neural network. The output of the RNN at time $t$ is computed using the state of the last hidden layer at $t$ as

$$y_t = \theta(W_{yh} h_t^n + b_y) \tag{2}$$

where $\theta$ is a non-linear operation such as sigmoid or hyperbolic tangent for binary classification or softmax for multiclass classification, $b_y$ is the bias term for the output layer and $n$ is the number of hidden layers in the architecture. The output of the RNN at desired time steps can then be used to compute the error and the network weights are updated based on the gradients computed using Back-propagation Through Time (BPTT). In simple RNNs, the function

$g$ is computed as a linear transformation of the input and previous hidden state, followed by an element wise non-linearity.

$$g(x_t, h_{t-1}) = \theta(W_{hx}x_t + W_{hh}h_{t-1} + b_h) \tag{3}$$

Such simple RNNs, however, suffer from the vanishing and exploding gradient problem [9]. To address this issue, a novel form of recurrent neural networks called the Long Short Term Memory (LSTM) networks were introduced in [9]. The key difference between simple RNNs and LSTMs is in the computation of $g$, which is done in the latter using a memory block. An LSTM memory block consists of a memory cell $c$ and three multiplicative gates which regulate the state of the cell - forget gate $f$, input gate $i$ and output gate $o$. The memory cell encodes the knowledge of the inputs that have been observed up to that time step. The forget gate controls whether the old information should be retained or forgotten. The input gate regulates whether new information should be added to the cell state while the output gate controls which parts of the new cell state to output. Like simple RNNs, LSTM networks can be made deep by stacking memory blocks. The output layer of the LSTM network can then be computed using Eq. 2. We refer the reader to [9] for more technical details on LSTMs.

### 3.2 Temporal Component of GeThR-Net

In this subsection, we describe the details of the temporal component, which is a temporally hybrid LSTM network that models modality specific temporal dynamics as well as the multimodal temporal dynamics. This network has three layers. The first layer models the modality specific temporal information using individual LSTM layers. Multimodal information do not interact with each other in this layer. In the second layer, the hidden representations from all the modalities are combined using a linear function, followed by sigmoid non-linearity, to create a single multimodal feature representation corresponding to each time step. Finally, in the third layer, a LSTM network is fed with the learned multimodal features from the second layer. The output from the third layer is fed into a softmax layer for estimating the classification confidence scores corresponding to each label. This component is fully trained in an end-to-end manner and does not require any explicit feature fusion modelling.

Now, we describe the technical details of these layers. We assume that there are total $M$ different modalities and total $T$ time-steps. The feature representation for modality $m$ corresponding to time instant $t$ is given by: $x_t^m$. Now, we describe the mathematical details:

- **First Layer:** The input to this layer is $x_t^m$ for modality $m$ at time instant $t$. If $LSTM_m$ denotes the LSTM layer for modality $m$ and if $h_t^m$ denotes the corresponding hidden representation at time $t$, then:

$$h_t^m = LSTM_m(x_t^m)$$

- **Second Layer:** In this layer, the hidden representations are combined using a linear function followed by a sigmoid non-linearity. The objective of using this layer is to combine features from multiple temporal modalities. Let us assume that $z_t$ denotes the concatenated hidden representation from all the modalities at time-step $t$. $W_z$ (same for all time-step $t$) denotes the weight matrix which combines the multimodal features and creates a representation $p_t$ at time instant $t$. $b_z$ denotes a linear bias and $\sigma$ is the sigmoid function.

$$z_t = (h_t^1, \cdots, h_t^m), \quad p_t = \sigma(W_z z_t + b_z)$$

- **Third Layer:** In this layer, one modality-independent LSTM layer is used to model the overall temporal dynamics of the multimodal feature representation $p_t$. Suppose, $LSTM_c$ denotes the combined LSTM and $h_t^c$ denotes the hidden representation from this LSTM layer at time $t$. $W_o$ is the weight matrix that linearly transforms the hidden representation. The output is propagated through a softmax function $\theta$ to obtain the final classification confidence values $y_t^c$ at time $t$. $b_o$ is a linear bias vector.

$$h_t^c = LSTM_c(p_t), \quad y_t^c = \theta(W_o h_t^c + b_o)$$

### 3.3 Non-temporal Component of GeThR-Net

Although it is important to model the temporal information in multimodal signals for accurate classification or any other tasks, often in real world scenarios multimodal information contains significant amount of noise and large intra-class variation along the temporal dimension. For example, videos of the activity 'cooking' often contain action segments such as 'changing thermostat' or 'drinking water' which are no way related to the actual label of the video. In those cases, modelling only the long-term temporal information in the video could lead to inaccurate results. Hence, it is important that we allow the proposed deep network to learn the non-temporal features too. We analyze videos from multiple datasets and observe that a simple classifier which is trained on 'frame-level' features (definition of frame could vary according to the features) could give a reasonable accuracy, especially when videos contain unrelated temporal segments. Please refer to Sect. 4.5 for more experimental results on this. Since our objective is to propose a generic deep network that could work with any kind of multimodal information, we add additional components to the GeThR-Net, which explicitly model the modality specific non-temporal information.

During training, for each modality $m$, we train a classifier where the set $\{x_t^m\}$, $\forall t$ is used as the training examples corresponding to the class of the multimodal signal. While testing for a given sequence, the predictions across all the time-steps are averaged to obtain the classifier confidence scores corresponding to all of the classes. In this paper, we have explored four different modalities: appearance, short-term motion, audio (spectrogram and MFCC) and skeleton. For appearance, short-term motion and audio-spectrogram, we use fine-tuned CNNs and for audio-MFCC and skeleton, we use SVMs as the non-temporal classifiers.

### 3.4 Combination

There are total $M + 1$ components in GeThR-Net, where the first one is the temporally hybrid LSTM network and the rest $M$ are the non-temporal modality specific classifiers corresponding to each modality. Once we independently train these $M + 1$ classifiers, their prediction scores are combined and a single class-label for each multimodal temporal sequence is predicted. We use a validation dataset to determine the relevant weights corresponding to each of the $M + 1$ components.

## 4 Experiments

Our goal is to demonstrate that the proposed GeThR-Net can be effectively applied to any kind of multimodal fusion. To achieve that, we perform thorough experimental evaluation and provide the details of the experimental results in this section.

### 4.1 Dataset Details

The dataset details are provided in this subsection.

**UCF-101 [28]:** UCF-101 is an action recognition dataset containing realistic action videos from YouTube. The dataset has 13,320 videos annotated into 101 different action classes. The average length of the video in this dataset is 6–7 sec. The dataset possess various challenges and diversity in terms of large variations in camera motion, object appearance and pose, cluttered background, illumination, viewpoint, etc. We evaluate the performance on this dataset following the standard protocol [28,40] by reporting the mean classification accuracy across three training and testing splits. We use the appearance and short-term motion modality for this dataset [24,40].

**CCV [13]:** The Columbia Consumer Videos (CCV) has 9,317 YouTube videos distributed over 20 different semantic categories. The dataset has events like 'baseball', 'parade', 'birthday', 'wedding ceremony', scenes like 'beach', 'playground', etc. and objects like 'cat', 'dog' etc. The average length of the video in this dataset is 80 sec long. For our experiments, we have used 7751 videos (3851 for training and 3900 for testing) as the remaining videos are not available on YouTube presently. In this dataset, the performance is measured by average precision (AP) for each class and the overall measure is given by mAP (mean average precision over 20 categories). In this dataset, we use three different modalities, i.e., appearance, short-term motion and audio.

**Multimodal Gesture Dataset [5] (MMG):** ChaLearn-2013 multimodal gesture recognition dataset is a large video database of 13,858 gestures from a lexicon of 20 Italian gesture categories. The focus of the dataset is on user independent multiple gesture learning. The dataset has RGB and depth images of the videos, user masks, skeletal model, and the audio information (utterance of the corresponding gesture by the actor), which are synchronous with the gestures performed. The dataset has 393 training, 287 testing, and 276 testing sequences. Each sequence is of duration between 1–2 min and contains 8–20 gestures. Furthermore, the test sequences also have 'distracter' (out of vocabulary) gestures apart from the 20 main gesture categories. For this dataset, we use the audio and skeleton modality for fusion because some of the top-performing methods [5] on this dataset also used these two modalities. The loose temporal boundaries of the gestures in the sequence is available during training and validation phase, however, at the time of testing, the goal is to also predict the correct order of gestures within the sequence along with the gesture labels. The final evaluation is defined in terms of edit distance (insertion, deletion, or substitution) between the ground truth sequence of labels and the predicted sequence of labels. The overall score is the sum of edit distance for all testing videos, divided by the total number of gestures in all the testing videos [5].

### 4.2 Modality Specific Feature Extraction

In this section, we describe the feature extraction method for different modalities - appearance, short-term motion, audio, and skeleton, which are used in this paper across three different datasets.

- **Appearance Features:** We adopted the VGG-16 [25] architecture to extract the appearance features. In this architecture, we change the number of neurons in fc7 layer from 4096 to 1024 to get a compressed lower dimensional representation of an input. We finetune the final three fully connected layers (fc6, fc7, and fc8) of the network pretrained on ImageNet using the frames of the training videos. The activations of the fc7 layer are taken as the visual representation of the frame provided as an input. While finetuning, we use minibatch stochastic descent with a fixed momentum of 0.9. The input size of the frame to our model is 224 × 224 × 3. Simple data augmentations are also done such as cropping and mirroring [11]. We adopt a dropout ratio of 0.5. The initial learning rate is set to 0.001 for fc6, and 0.01 for fc7 and fc8 layers as the weights of last two layers are learned from scratch. The learning rate is reduced by factor of 10 after every 10,000 iterations.
- **Short-Term Motion Features:** To extract the features, we adopted the method proposed in the recent two-stream CNN paper [24]. This method stacks the optical flows computed between pairs of adjacent frames over a time window and provides it as an input to CNN. We used the same VGG-16 architecture (as above) with 1024 neurons in fc7 layer, and pre-training on ImageNet for the extraction of short-term motion features. However, unlike the previous case (where input to the model was an RGB image comprising of three channels), the input to this network is a 10-frame stacking of optical flow fields (x and y direction), and thus the convolution filters in the first layer are different from those of the appearance network. We adopt a high dropout rate of 0.8 and set the initial learning rate to 0.001 for all the layers. The learning rate is reduced by a factor of 10 after every 10,000 iterations.

- **Audio Features:** We use two different kinds of feature extraction method for audio modality.
  - **Spectrogram Features:** In this method, we extract the spectrogram features from audio signal using a convolutional neural network [21]. We divide the video into multiple overlapping 1 sec clips and then, apply the Short Time Fourier Transformation to convert each one second 1-d audio signal into a 2-D image (namely log-compressed mel-spectrograms with 128 components) with the horizontal axis and vertical axis being time-scale and frequency-scale respectively. The features are extracted from these spectrogram images by providing them as input to a CNN. In this case, we use AlexNet [14] architecture and the network was pre-trained on ImageNet. We finetune the final three layers of network with respect to the spectrogram images of training videos to learn the 'spectrogram-discriminative' CNN features. We also change the number of nodes in fc7 layer to 1024 and use the activations of fc7 layer as the representation of a spectrogram image. The learning rate and dropout parameters are same as mentioned in the appearance feature extraction case.
  - **MFCC Features:** We use MFCC features for the MMG dataset. The spectrogram based CNN features were not used for this dataset as the temporal extent of each gesture was very less (1–2 sec), making it difficult to extract multiple spectrograms along the temporal dimension. In this method, speech signal of a gesture was analyzed using a 20ms Hamming window with a fixed frame rate of 10ms. Our feature consists of 12 Mel Frequency Cepstral Coefficients (MFCCs) along with the log energy ($MFCC_0$) and their first and second order delta values to capture the spectral variation. We concatenated 5 adjacent frames together in order to adhere to the 20 fps of videos in the MMG dataset. Hence, we have a feature of dimension of $39 \times 5 = 195$ for each frame of the video. The data was also normalized such that each of the features (coefficients, energy and derivatives) extracted have zero mean and one variance.
- **Skeleton Features:** We use the skeleton features for the MMG dataset. We employ the feature extraction method proposed in [36,43] to characterize the action information which includes the posture feature, motion feature and offset feature. Out of 20 skeleton joint locations, we use only 9 upper body joints as they are the most discriminative for recognizing gestures.

### 4.3 Methods Compared

To establish the efficacy of the proposed approach, we compare GeThR-Net with several baselines. The baselines were carefully designed to cover several temporal and non-temporal feature fusion methods. We provide the architectural details of these baselines in Fig. 2 for easy understanding of their differences.

(a) **NonTemporal-M:** In this baseline, we train modality specific non-temporal models and predict label of a temporal sequence based on the average over

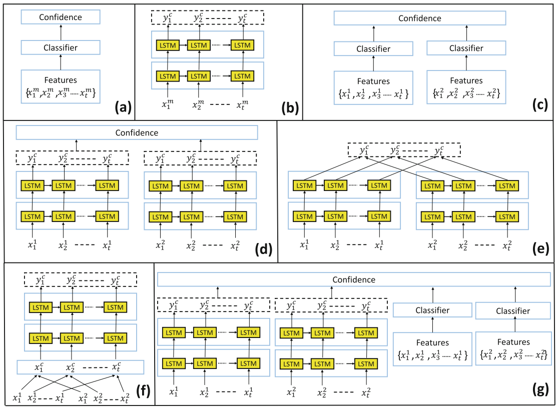

**Fig. 2.** Different baselines which are compared with GeThR-Net. (a) NonTemporal-M. (b) Temporal-M. (c) NonTemporal-AM (d) Temporal-AM (late fusion) (e) Temporal-EtoE-AM (late fusion) (f) Temporal-AM (early fusion) (g) Temporal-AM+NonTemporal-AM.

all predictions across time. For appearance, short-term motion and audio spectrogram, we use CNN features (Sect. 4.2) followed by a softmax layer for classification. For audio MFCC and Skeleton, we use the features extracted using the methods described in Sect. 4.2 followed by SVM classification. Multimodal fusion is not performed for label prediction in these baselines.

(b) **Temporal-M:** For this baseline, we feed the modality specific features (as described in the last subsection), to LSTM networks for the temporal modelling and label prediction. Here also, features from multiple modalities are not fused for classification.

(c) **NonTemporal-AM (all modality combined):** In this baseline, the outputs from the modality specific non-temporal baselines (CNN/SVM) are linearly combined for classification. The combination weights are automatically learned from validation datasets.

(d) **Temporal-AM (late fusion, all modality combined):** Here also, the outputs from the modality specific temporal baselines (LSTMs) are linearly combined for classification. This is a late fusion approach.

(e) **TemporalEtoE-AM (late fusion, all modality combined):** In this baseline, we add a linear layer on top of the modality specific temporal baselines and use an end-to-end training approach for learning the weights of the combination layer. This is also a late fusion approach.

(f) **Temporal-AM (early fusion, all modality combined):** Features from multiple modalities are linearly combined and then forward propagated through a LSTM for classification. This is an early fusion approach.

(g) **Temporal-AM+NonTemporal-AM (all modality combined):** In this baseline, the outputs from all the modality specific temporal and nontemporal baselines are combined for the final label prediction. Here also, we use a validation dataset for predicting the optimal weights corresponding to each of these components.

(h) **TemporallyHybrid-AM (proposed, all modality combined):** This method uses only the temporally hybrid component of the proposed approach. The non-temporal components' outputs are not used. This network is completely trained in an end-to-end fashion (See the temporal component in Fig. 1).

(i) **GeThR-Net:** This is the proposed approach (See Fig. 1).

### 4.4 Implementation Details

We used the initial learning rate of 0.0002 for all LSTM networks. It is reduced by a factor of 0.9 for every epoch starting from the 6-th epoch. We set the dropout rate at 0.3. For the baseline methods of temporal modelling, Temporal-M, Temporal-AM and TemporalEtoE-AM, we tried different combinations for the number of hidden layers and the number of units in each layer and chose the one which led to the optimal performance on the validation set. Since, the feature dimension is high (1024) in UCF-101 and CCV dataset, the number of units in each layer is varied from 256 to 768 in the intervals of 32. While in case of MMG, it is varied from 64 to 512 in the same interval. The number of layers in the baselines were varied between 1 and 3 for all of the datasets.

For the proposed temporally hybrid network (TemporallyHybrid-AM) component also, the number of units in the First-layer LSTM corresponding to each modality, the number units in the linear Second-layer and the number of units in Third-layer multimodal LSTM are chosen based upon the performance on the validation dataset. For UCF-101 dataset, the First-layer has 576 units for both the appearance and short-term modality. The Second-layer has 768 units and the Third-layer has 448 units. For CCV dataset, all the three modalities, appearance, short-term motion and audio have 512 units in the First-layer. In CCV, the Second-layer has 896 units and the Third-layer has 640 units. For MMG dataset, the First-layer has 256 units for skeleton modality and 192 units for audio modality. The Second-layer has 384 units and the Third-layer has 256 units. Note that these parameters differ across the datasets due to the variation in the input feature size and the inherent complexity of the datasets.

### 4.5 Discussion on Results

In this section, we compare GeThR-Net with various baseline methods (Sect. 4.3) and several recent state-of-the-art methods on three different datasets. The results corresponding to all the baselines and the proposed approach are summarized in Table 1. In the first two slabs of the table, results from individual modalities are shown using the temporal and non-temporal components. In the next three slabs, results for different fusion strategies across modalities are shown for both the temporal and non-temporal components. In the final slab of the table, results obtained from the proposed temporally hybrid component and GeThR-Net are shown.

**Table 1.** Comparison of GeThR-Net with baseline methods on UCF-101, CCV and Multimodal Gesture recognition (MMG) dataset. UCF-101: M1 is appearance, M2 is short-term motion and classification accuracy is reported. CCV: M1 is appearance, M2 is short-term motion, M3 is audio and mean average precision (mAP) is reported. MMG: M1 is audio, M2 is skeleton and normalized edit distance is reported.

| Dataset | Modalities Used |
|---|---|
| UCF-101 | Appearance (M1) |
|  | Short term Motion (M2) |
| CCV | Appearance (M1) |
|  | Short-term Motion (M2) |
|  | Audio (M3) |
| MMG | Audio (M1) |
|  | Skeleton (M2) |

| Methods | UCF-101 (Accuracy) | CCV (mAP) | MMG (edit) |
|---|---|---|---|
| NonTemporal-M1 | 76.3 | 76.7 | 0.988 |
| NonTemporal-M2 | 86.8 | 57.3 | 0.782 |
| NonTemporal-M3 | - | 30.3 | - |
| Temporal-M1 | 76.6 | 71.7 | 0.284 |
| Temporal-M2 | 85.5 | 55.1 | 0.361 |
| Temporal-M3 | - | 28.5 | - |
| NonTemporal-AM | 89.9 | 78.5 | 0.776 |
| Temporal-AM (late fusion) | 88.0 | 75.0 | 0.156 |
| TemporalEtoE-AM (late fusion) | 88.4 | 72.5 | 0.155 |
| Temporal-AM (early fusion) | 86.5 | 73.1 | 0.190 |
| Temporal-AM + NonTemporal-AM | 90.2 | 79.2 | 0.155 |
| TemporallyHybrid-AM | 89.0 | 74.0 | **0.152** |
| GeThR-Net | **91.1** | **79.3** | **0.152** |

- **UCF-101 [28]:** For UCF-101, we report the test video classification accuracy. GeThR-Net achieves an absolute improvement of 3.1 %, 2.7 % and 4.6 % over Temporal-AM (late fusion), TemporalEtoE-AM (late fusion) and Temporal-AM (early fusion) baselines respectively. This empirically shows that the proposed approach is significantly better in capturing the complementary temporal aspects of different modalities compared to the late and early fusion based methods. GeThR-Net also gives an absolute improvement of 0.9 % over a strong baseline method of combining temporal and non-temporal aspects of different modalities (Temporal-AM+Non-Temporal-AM). This further establishes the efficacy of the proposed architecture. We also compare the results produced by GeThR-Net with several recent papers which reported results on UCF-101 (see Table 2). Out of the seven approaches we compare, we are better than five of them and comparable to two [34,40] of them. As pointed out earlier, the goal of this paper is to develop a general deep learning framework which can be used for multimodal fusion in different kinds of tasks. The results on UCF-101 clearly shows that GeThR-Net can be effectively used for the short action recognition task (average duration 6–7 seconds).

- **CCV [5]:** We also perform experiments on the CCV dataset to show that GeThR-Net can also be used for longer action recognition (average duration 80 seconds). In this dataset, we report the mean average precision (in a scale of 0–100) for all the algorithms which we compare. In CCV also, GeThR-Net is better than Temporal-AM (late fusion), TemporalEtoE-AM (late fusion) and Temporal-AM (early fusion) baselines by an absolute mAP of 4.3, 6.8 and 6.2 respectively. However, GeThR-Net performs comparable (mAP of 79.3 compared to 79.2) to a strong baseline method of combining temporal and non-temporal aspects of different modalities (Temporal-AM+Non-Temporal-AM). We also wanted to compare GeThR-Net with several recent approaches which also reported results on the CCV dataset. However, a fair comparison was not possible because several videos from CCV were unavailable from youtube. We used only 7,751 videos for training and testing as opposed to 9,317 videos in the original dataset. In spite of that, to get an approximate idea about how GeThR-Net performs compared to these methods, we provide some comparisons. The mAP reported on CCV by some of the recent methods are: 70.6 [39], 64.0 [45], 63.4 [16], 60.3 [42], 68.2 [15], 64.0 [10] and 83.5 [40]. We perform better (mAP of 79.3) than six of these methods.
- **MMG [5]:** In this dataset, we report the normalized edit distance (lower is better) [5] corresponding to each method. The normalized edit distance obtained by GeThR-Net is lower than the other multimodal baselines such as Temporal-AM (late fusion), TemporalEtoE-AM (early fusion), Temporal-AM (late fusion) and Temporal-AM+NonTemporal-AM by 0.004, 0.003, 0.038 and 0.003 respectively. We are also significantly better than modality specific temporal baselines, e.g.: GeThR-Net gives a normalized edit distance of only 0.152 compared to 0.284 and 0.361 produced by Temporal-M1 (audio) and Temporal-M2 (skeleton) respectively. The results on this dataset demonstrates that GeThR-Net performs well in fusing multimodal information from audio-MFCC and skeleton. The edit distance obtained from GeThR-Net is one of the top-three edits distances reported in the Chalearn-2013 multimodal gesture recognition competition [5].

**Table 2.** Comparison of GeThR-Net with state-of-the-art methods on UCF-101.

| IDT + FV [32] | IDT + HSV [23] | Two-stream [24] | LSTM [19] | TDD + FV [33] | Two-stream2 [34] | Fusion [40] | GeThR-Net |
|---|---|---|---|---|---|---|---|
| 85.9 | 87.9 | 88.0 | 88.6 | 90.3 | 91.4 | 91.3 | 91.1 |

From the results on these datasets, it is clear that GeThr-Net is effective in fusing different kinds of multimodal information and also applicable to different end-tasks such as short action recognition, long action recognition and gesture recognition. That empirically shows the generalizability of the proposed deep network.

## 5 Conclusion

In this paper, we propose a novel deep neural network called GeThR-Net for multimodal temporal information fusion. GeThR-Net has a temporally hybrid recurrent neural network component that models modality specific temporal dynamics as well as the temporal dynamics in a multimodal feature space. The other components in the GeThR-Net are used to capture the non-temporal information. We perform experiments on three different action and gesture recognition datasets and show that GeThR-Net performs well for any general multimodal fusion task. The experimental results are performed on four different modalities with maximum three modality fusion at a time. However, GeThR-Net can be used for any kind of modality fusion without any upper bound on the number of modalities that can be combined.

## References

1. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia Syst. **16**(6), 345–379 (2010)
2. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Sequential deep learning for human action recognition. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 29–39. Springer, Heidelberg (2011)
3. Chen, S., Jin, Q.: Multi-modal dimensional emotion recognition using recurrent neural networks. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, pp. 49–56. ACM (2015)
4. Chen, X., C., L.Z.: Mind's eye: a recurrent visual representation for image caption generation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
5. Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Lopes, O., Guyon, I., Athitsos, V., Escalante, H.: Multi-modal gesture recognition challenge 2013: dataset and results. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 445–452. ACM (2013)
6. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: ICML (2014)
7. Graves, A., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks. In: ICASSP. IEEE (2013)
8. Graves, A., Schmidhuber, J.: Offline handwriting recognition with multidimensional recurrent neural networks. In: NIPS (2009)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
10. Jhuo, I.H., Ye, G., Gao, S., Liu, D., Jiang, Y.G., Lee, D.T., Chang, S.F.: Discovering joint audio-visual codewords for video event detection. Mach. Vis. Appl. **25**(1), 33–47 (2014)
11. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
12. Jiang, W., Cotton, C., Chang, S.F., Ellis, D., Loui, A.: Short-term audio-visual atoms for generic video concept classification. In: Proceedings of the 17th ACM International Conference on Multimedia, pp. 5–14. ACM (2009)

13. Jiang, Y.G., Ye, G., Chang, S.F., Ellis, D., Loui, A.C.: Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, p. 29. ACM (2011)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
15. Liu, D., Lai, K.T., Ye, G., Chen, M.S., Chang, S.F.: Sample-specific late fusion for visual category recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
16. Ma, A.J., Yuen, P.C.: Reduced analytic dependency modeling: robust fusion for visual recognition. Int. J. Comput. Vis. **109**, 233–251 (2014)
17. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). In: ICLR (2015)
18. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.L.: Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090 (2014)
19. Ng, J.Y., Hausknecht, M.J., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: deep networks for video classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June, 2015
20. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pp. 689–696 (2011)
21. Van den Oord, A., Dieleman, S., Schrauwen, B.: Deep content-based music recommendation. In: Advances in Neural Information Processing Systems, pp. 2643–2651 (2013)
22. Paleari, M., Lisetti, C.L.: Toward multimodal fusion of affective cues. In: Proceedings of the 1st ACM International Workshop on Human-Centered Multimedia, pp. 99–108. ACM (2006)
23. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. CoRR (2014)
24. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS, pp. 568–576 (2014)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. Snoek, C.G., Worring, M., Smeulders, A.W.: Early versus late fusion in semantic video analysis. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 399–402. ACM (2005)
27. Sohn, K., Shang, W., Lee, H.: Improved multimodal deep learning with variation of information. In: Advances in Neural Information Processing Systems, pp. 2141–2149 (2014)
28. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
29. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: Advances in Neural Information Processing Systems, pp. 2222–2230 (2012)
30. Stein, B.E., Stanford, T.R., Rowland, B.A.: The neural basis of multisensory integration in the midbrain: its organization and maturation. Hear. Res. **258**(1), 4–15 (2009)
31. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR (2015)

32. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: The IEEE International Conference on Computer Vision (ICCV), December 2013
33. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors. In: CVPR, pp. 4305–4314 (2015)
34. Wang, L., Xiong, Y., Wang, Z., Qiao, Y.: Towards good practices for very deep two-stream convnets. CoRR (2015)
35. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural Comput. **1**, 263–269 (1989)
36. Wu, D., Shao, L.: Multimodal dynamic networks for gesture recognition. In: Proceedings of the ACM International Conference on Multimedia, pp. 945–948. ACM (2014)
37. Wu, Y., Chang, E.Y., Chang, K.C.C., Smith, J.R.: Optimal multimodal fusion for multimedia data analysis. In: Proceedings of the 12th Annual ACM International Conference on Multimedia, pp. 572–579. ACM (2004)
38. Wu, Z., Jiang, Y.G., Wang, J., Pu, J., Xue, X.: Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In: Proceedings of the ACM International Conference on Multimedia, pp. 167–176. ACM (2014)
39. Wu, Z., Jiang, Y.G., Wang, J., Pu, J., Xue, X.: Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In: Proceedings of the 22nd ACM International Conference on Multimedia, MM 2014 (2014)
40. Wu, Z., Wang, X., Jiang, Y.G., Ye, H., Xue, X.: Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In: Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, pp. 461–470. ACM (2015)
41. Xie, Z., Guan, L.: Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools. In: 2013 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2013)
42. Xu, Z., Yang, Y., Tsang, I., Sebe, N., Hauptmann, A.G.: Feature weighting via optimal thresholding for video analysis. In: 2013 IEEE International Conference on Computer Vision (ICCV) (2013)
43. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 14–19. IEEE (2012)
44. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4507–4515 (2015)
45. Ye, G., Liu, D., Jhuo, I.H., Chang, S.F., et al.: Robust late fusion with rank minimization. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3021–3028. IEEE (2012)