HAL
open science

Temporal Multimodal Fusion for Video Emotion Classification in the Wild

Valentin Vielzeuf, Stéphane Pateux, Frédéric Jurie

To cite this version:

Valentin Vielzeuf, Stéphane Pateux, Frédéric Jurie. Temporal Multimodal Fusion for Video Emotion Classification in the Wild. ACM - ICMI 2017, Nov 2017, Glasgow, United Kingdom. hal-01590608

HAL Id: hal-01590608

https://hal.science/hal-01590608v1

Submitted on 20 Sep 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Temporal Multimodal Fusion for Video Emotion Classification in the Wild

Valentin Vielzeuf*
Orange Labs
Cesson-Sévigné, France
valentin.vielzeuf@orange.com

Stéphane Pateux
Orange Labs
Cesson-Sévigné, France
stephane.pateux@orange.com

Frédéric Jurie
Normandie Univ., UNICAEN, ENSICAEN, CNRS
Caen, France
frederic.jurie@unicaen.fr

ABSTRACT

This paper addresses the question of emotion classification. The task consists in predicting emotion labels (taken among a set of possible labels) best describing the emotions contained in short video clips. Building on a standard framework – lying in describing videos by audio and visual features used by a supervised classifier to infer the labels – this paper investigates several novel directions. First of all, improved face descriptors based on 2D and 3D Convolutional Neural Networks are proposed. Second, the paper explores several fusion methods, temporal and multimodal, including a novel hierarchical method combining features and scores. In addition, we carefully reviewed the different stages of the pipeline and designed a CNN architecture adapted to the task; this is important as the size of the training set is small compared to the difficulty of the problem, making generalization difficult. The so-obtained model ranked 4th at the 2017 Emotion in the Wild challenge with the accuracy of 58.8 %.

KEYWORDS

Emotion Recognition; Multimodal Fusion; Recurrent Neural Network; Deep Learning;

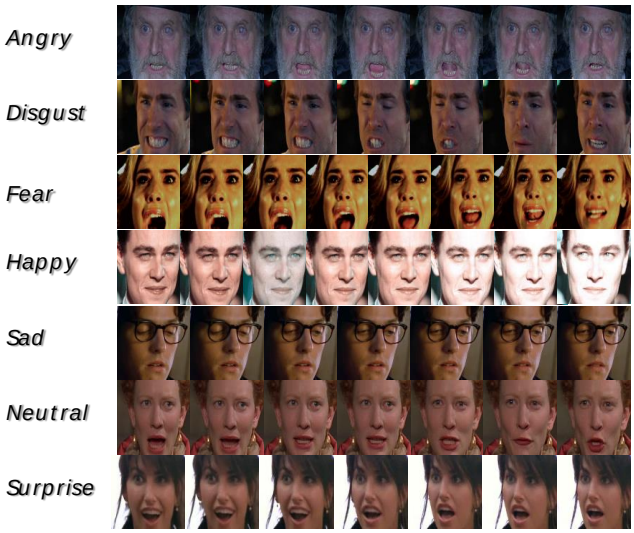

Figure 1: The 7 emotions of the AFEW dataset: each row represents an emotion with a set of faces sampled across a representative video clip. Please note that if most of the video clips do contain faces, some of them don't.

1 INTRODUCTION AND RELATED WORK

Emotion recognition is a topic of broad and current interest, useful for many applications such as advertising [27] or psychological disorders understanding [35]. It is also a topic of importance for research in other areas *e.g.*, video summarization [37] or face normalization (expression removal). Even if emotion recognition could appear to be an almost solved problem in laboratory controlled conditions, there are still many challenging issues in the case of videos recorded in the wild.

This paper focuses on the task of emotion classification in which each video clip has to be assigned to one and only one emotion, based on its audio/video content. The classes are usually the six basic emotions *i.e.*, anger, disgust, fear, happiness, sadness and surprise, in addition to the neutral class, as in the *Audio-video Emotion Recognition* sub-task of the Emotion Recognition in the Wild Challenge [12]. More precisely, this paper presents our methodology, experiments as well as the results we obtained in the 2017 edition of the Emotion Recognition in the Wild Challenge [10].

Emotion recognition has received a lot of attention in the scientific literature. One large part of this literature deals with the possible options for defining and representing emotions. If the use of discrete classes [31] such as joy, fear, angriness, *etc.* is the most straightforward way to do it, it can be more interesting to represent emotions by their degrees of arousal and valence, as proposed in [6]. In the restricted case of facial expressions, action units can also be used, focusing on the activation of different parts of the face [15]. Links can be made between these two representations: the discrete classes can be mapped into the arousal valence space [28] and can be deduced from the action units [24].

Another important part of the literature focuses on the ways to represent audio/video contents by features that can be subsequently used by classifiers. Early papers make use of (i) hand-crafted features such as Local Binary Patterns (LBP), Gabor features, discrete cosine transform for representing images, and Linear Predictive Coding coefficients (LPC), relative spectral transform - linear perceptual prediction (RASTA-PLP), modulation spectrum (ModSpec) or Enhanced AutoCorrelation (EAC) for audio, (ii) standard classifiers such as SVN or KNN for classification (see [22] for details). [38], the winner of EmotiW'15, demonstrated the relevance of Action Units for the recognition of emotions. [3] was among the first to propose to learn the features instead of using hand-crafted descriptors, relying on Deep Convolutional Networks. More recently, [18], the winner of EmotiW'16, has introduced the C3D feature which is an efficient spatio-temporal representation of faces.

*Also with Normandie Univ., UNICAEN, ENSICAEN, CNRS.

The literature on emotion recognition from audio/video contents also addresses the question of the fusion of the different modalities. A modality can be seen as one of the signals allowing to perceive the emotion. Among the recent methods for fusing several modalities, we can mention the use of two-stream ConvNets [32], ModDrop [29] or Multiple Kernel Fusion [9]. The most used modalities, in the context of emotion recognition, are face images and audio/speech, even if context seems also to be of tremendous importance [5]. For instance, the general understanding of the scene, even based on simple features describing the whole image, may help to discriminate between two candidate classes.

As most of the recent methods for emotion recognition are supervised and hence requires some training data, the availability of such resources is becoming more and more critical. Several challenges have collected useful data. The AVEC challenge [34] focuses on the use of several modalities to track arousal and valence in videos recorded in controlled conditions. The Emotionet challenge [7] proposes a dataset of one million images annotated with action units and partially with discrete compound emotion classes. Finally, the Emotion in the Wild challenge [11] deals with the classification of short video clips into seven discrete classes. The videos are extracted from movies and TV shows recorded "in the wild". The ability to work with data recorded in realistic situations, including occlusions, poor illumination conditions, presence of several people or even scene breaks is indeed very important.

As aforementioned, this paper deals with our participation to the Emotion In the Wild 2017 (EmotiW) challenge. We build on the state-of-the-art pipeline of [18] in which i) audio features are extracted with the OpenSmile toolkit [16], ii) two video features are computed, one by the C3D descriptor, the other by the VGG16-Face model [30] fine-tuned with FER2013 face emotion database and introduced into a LSTM network. Each one of these 3 features is processed by its own classifier, the output of the 3 classifiers being combined through late fusion. Starting from this pipeline, we propose to improve it in 3 different ways, which are the main contributions of our approach.

First, the recent literature suggests that late fusion might not be the optimal way to combine the different modalities (see *e.g.*, [29]). This paper investigates different directions, including an original hierarchical approach allowing to combine scores (late fusion) and features (early fusion) at different levels. It can be seen as a way to combine information at its optimal level of description (from features to scores). This representation addresses an important issue of fusion, which is to ensure the preservation of unimodal information while being able to exploit cross-modal information.

Second, we investigate several ways to better use the temporal information in the visual descriptors. Among several contributions, we propose a novel descriptor combining C3D and LSTM.

Third, it can be observed that the amount of training data (773 labeled short video clips) is rather small compared to the number of the parameters of standard deep models, considering the complexity and diversity of emotions. In this context, supervised methods are prone to over-fitting. We show in the paper how the effect of over-fitting can be reduced by carefully choosing the number of parameters of the model and favoring transfer learning whenever it is possible.

Figure 2: Comparison of the bounding boxes given by our detection/normalization procedure (color images) with those provided with the AFEW dataset (gray-scale images), on 4 random faces of AFEW. Our bounding boxes are slightly larger.

The rest of the paper is organized as follows: a presentation of the proposed model is done in Section 2, detailing the different modalities and the fusion methods. Then Section 3 presents the experimental validation as well as the results obtained on the validation and test sets during the challenge.

2 PRESENTATION OF THE PROPOSED APPROACH

Figure 3 presents an overview of our approach, which is inspired by the one of [18]. Modalities we consider are extracted from audio and video, associated with faces analyzed with two different models (2D CNN, C3D). The two main contributions are the temporal fusion and the novel C3D/LSTM descriptor.

On overall, our method works as follows: on the one hand, the OpenSmile library [16] is used to produce 1582 dimensional features used by a two-layer perceptron to predict classes as well as compact descriptors (279-d vectors) from audio. On the other hand, video classification is based on face analysis. After detecting the faces and normalizing their appearance, one set of features is computed with the VGG-16 model [30] while another set of features is obtained by the C3D model [18]. In both cases, 4096-d features are produced. Temporal fusion of these so-obtained descriptions is done across the frames of the sequences, producing per modality scores and compact descriptors (304 dimensional vectors). The 3 modalities (audio, VGG-faces and C3D-faces) are then combined using both the score predictions and the compact representations.

After explaining how the faces are detected and normalized, the rest of the section gives more details on how each modality is processed, and how the fusion is performed.

2.1 Face Detection and Alignment

The EmotiW challenge provides face detections for each frame of each video. However, we preferred not to use these annotations but to detect the faces ourselves. The motivation is twofold. First, we want to be able to process any given video and not only those of EmotiW (*e.g.*, for adding external training data). Second, it is necessary to master the face alignment process for processing the

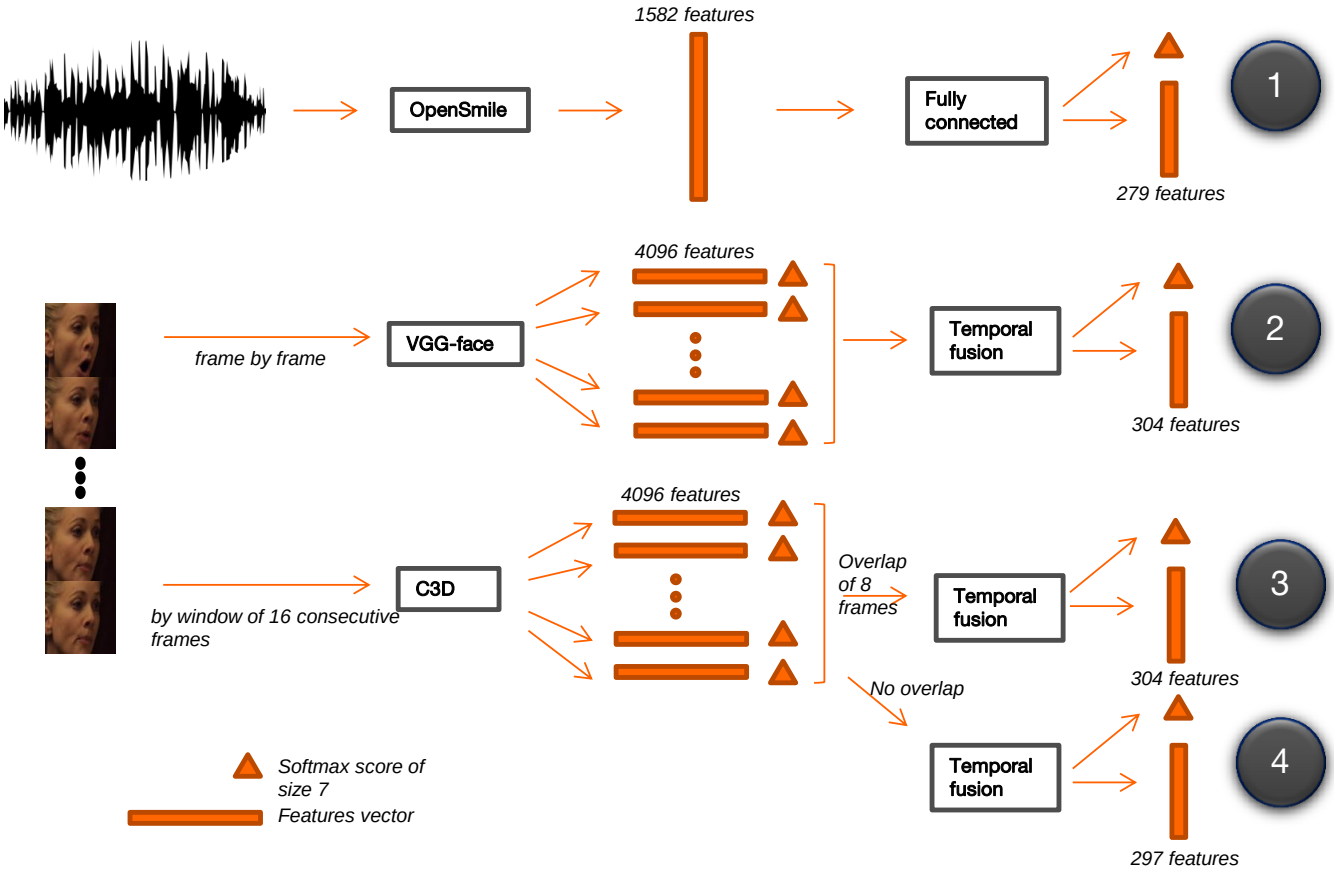

Figure 3: Our model includes audio, VGG-LSTM and C3D-LSTM (with or without overlapping) modalities. Each modality gives its own feature vectors and scores, which can be fused using different methods. It can be noted that the dimensions of the features vectors are chosen to have balanced contributions between modalities, making a trade off between the number of parameters and the performance.

images of face datasets when pre-training the VGG models (see Section 2.3 where FER dataset is used to pre-train the VGG model).

For face detection, we use an internal detector provided by Orange Labs. We found out from our observations that this face detector detects more faces (20597 versus 19845 on the validation set) while having a lower false positive rate (179 false positives versus 908 on the validation set) than the one use to provide EmotiW annotations.

This detector is applied frame per frame. If one or several faces are detected, a tracking based on their relative positions allows to define several face tracks. In a second time, an alignment based on the landmarks [26] is done. We choose the longest face sequence in each video and finally apply a temporal smoothing of the positions of the faces to filter out jittering.

Figure 2 compares one of our normalized faces (after detection / alignment) with the one provided with the challenge data.

2.2 Audio features and classifier

The audio channel of each video is fed into the OpenSmile toolkit [16][1], as most of the EmotiW competitors [11], to get a description vector of length 1582.

A commonly used approach for audio is then to learn a classifier on top of the OpenSmile features. Support Vector Machine [18] [38] seems to be the dominant choice for the classification, even if there are some other approaches like Random Forests [17].

To be able to control more finely the dimensionality of the description vector (the one used later on during the fusion process), we learn a two-layer Perceptron with reLu activation on the OpenSmile description, using batch normalisation and dropout. During inference, we extract for each video a description vector of size 279 – the hidden layer of the perceptron – along with the softmax score.

2.3 Representing Faces with 2D Convolutional Neural Network

A current popular method (see *e.g.*, [18, 23]) for representing faces in the context of emotion recognition, especially in the EmotiW

[1]We use the `emobase2010` configuration file.

challenge, is to fine-tune a pre-trained 2D CNN such as the VGG-face model [30] on the images of an emotion images dataset (*e.g.*, the FER 2013 dataset). Using a pre-trained model is a way to balance the relatively small size of the EmotiW dataset (AFEW dataset). The images of the FER 2013 dataset [20] are first processed by detecting and aligning the faces, following the procedure explained in Section 2.1. We then fine-tune the VGG-face model on FER 2013 dataset, using both the training and the public test set; during training we use data augmentation by jittering the scale, flipping and rotating the faces. The aim is to make the network more robust to small misalignment of the faces. We also apply a strong dropout on the last layer of the VGG (keeping only 5% of the nodes) to prevent over-fitting. We achieve a performance of 71.2% on the FER private test set, which is slightly higher than the previously published results [18, 20].

To assess the quality of the description given by this fine-tuned VGG model to emotion recognition, we first benchmarked it on the validation set of SFEW [13], which is a dataset containing frames of the videos of AFEW (those used by the EmotiW challenge). We achieved a score of 45.2% without retraining the model on SFEW, while the state-of-the-art result [25] is 52.5%, using a committee of deep models, and the challenge baseline is of 39.7% [13].

The face sequences detected in AFEW videos are resized to 224x224 and fed into the fine-tuned VGG-face model. We then extract the 4096 length fc6 layer, following the same pipeline as Fan et al. [17, 18]. We also compute the softmax score for each frame.

2.4 Representing Faces with 3D Convolutional Neural Network

3D convolutional neural networks have been shown to give good performance in the context of facial expressions recognition in video [2, 33]. Fan *et al.* [18] fine-tuned a pre-trained 'sport1m' model on randomly chosen windows of 16 consecutive faces. During inference, the model is applied to the central frame of the video.

One limitation of this approach is that, at test time, there is no guaranty that the best window for capturing the emotion is in the middle of the video. The same problem occurs during training: a large part of the windows (randomly) selected for training does not contain any emotion or does not contain the correct emotion. Indeed, videos are annotated as a whole but some frames can have different labels or, sometimes, no expression at all.

We address this limitation in the following way: we first fine-tune a C3D-sport1m model using all the windows of each video, optimizing the classification performance until the beginning of convergence. Then, to be able to learn more from the most meaningful windows than from the others, we weigh each window based on its scores. More precisely, for the ith video and the jth window of this video, at epoch t, the weight $w_{i,j}$ is computed as:

$$w_{i,j} = e^{\frac{-s_{i,j}}{T(t)}} \tag{1}$$

where $T(t)$ is a temperature parameter decreasing with epoch t and $s_{i,j}$, the score of the window j of the video i. We then normalize the weight to ensure that for each video i, $\sum_j w_{i,j} = 1$. A random grid search on the hyper-parameters, including the temperature descent, is made on the validation set.

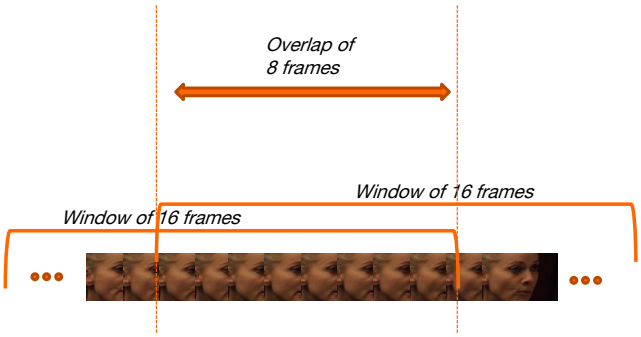

Figure 4: In the case of 8 frames overlap, we can see that the 16-frame windows are sharing half of their faces.

During inference, the face sequences detected in AFEW videos are resized to 112x112. We then split a video into several windows of 16 consecutive frames[2], with or without overlapping between the windows (as shown in Figure 4), and fed it into the weighted C3D. We then extract the second 4096-d fully connected layer and the last softmax layer of the C3D for each window of the videos.

The method we propose bears similarities with Multiple Instance Learning (MIL) [1]. Within the framework of MIL, each video can be considered as a bag of windows, with one single label. A straightforward way to apply MIL would be to train the C3D on each video as a bag of windows, and add a final layer for choosing the prediction as the one with the maximum score among all the scores of the batch. The loss would be then computed from this prediction. The weights defined in Eq.(1) play this role by selecting iteratively the best scoring windows.

2.5 Per Modality Temporal Fusion

Both VGG and weighted C3D representations are applied to each frame of the videos, turning the videos into temporal sequences of visual descriptors. We will name the elements of those sequences as "descriptors", whether it is the description of a frame or of a window.

To classify these sequences of descriptors, we investigated several methods. The most straightforward one is to score each descriptor and to take the maximum of the softmax scores as the final prediction. Similarly, the maximum of the means of the softmax across time can also be considered.

To better take into account temporal dependencies between consecutive descriptors, another option is to use Long Short-Term Memory recurrent neural networks (LSTM) [2, 19]. Unlike [17, 18], we chose to use a variable length LSTM allowing us to take all the descriptors as inputs.

To prevent over-fitting, we also applied dropout to the LSTM cells and weight decay to the final fully connected layer. A random grid search on hyper-parameters is then applied for each one of these models.

[2]The sequences with fewer than 16 frames were padded by themselves to reach a sufficient length.

Both VGG-LSTM and C3D-LSTM are used to give one description vector (output of the final fully connected layer of the LSTM) and one softmax score for each video.

The final VGG-LSTM architecture has 2230 hidden units for each LSTM-cell and a final fully connected layer with 297 hidden units. The maximal length of the input sequence is of 272 frames.

The final C3D-LSTM architecture has 1324 hidden units for each LSTM-cell and a final fully connected layer of 304 hidden units. The maximal length of the input sequence is of 34 windows (overlap of 8 frames) or 17 windows (no overlap).

2.6 Multimodal fusion

Last but not the least, the different modalities have to be efficiently combined to maximize the overall performance.

As explained in Section 1, two main fusion strategies can be used *i.e.* score fusion (late fusion), which consists in predicting the labels based on the predictions given by each modality, or features fusion (early fusion), which consists in taking as input the latent features vectors given by each modality and learning a classifier on the top of them.

During the last editions of the challenge, most of the papers focused on score fusion, using SVM [4], Multiple Kernel Fusion [9] or weighted means [18]. Differently, several authors tried to train audio and image modalities together (see *e.g.*, [8]), combining early features and using soft attention mechanism, but they didn't achieve state-of-the-art performance. We propose an approach combining both.

The rest of the section describes the four different fusion methods we experimented with, including a novel method (denominated as *score trees*).

Baseline score fusion. We experimented with several standard score fusion, like majority voting, means of the scores, maximum of the scores and linear SVM.

Fully connected neural network with modality drop. We also experimented with the ModDrop method of Neverova *et al.* [29]. It consists in applying dropout by modality, learning the cross-modality correlations while keeping unimodal information. [29] reports state-of-the-art results on gesture recognition. We apply this method to our audio, C3D-LSTM and VGG-LSTM features, as shown in Figure 5. According to [29], this is much better than simply feeding the concatenation of the modalities features into a fully connected neural network and letting the network learn a joint representation. Indeed, the fully connected model would be unable to preserve unimodal information during cross-modality learning.

An important step to make convergence possible with ModDrop is to first learn the fusion without cross-modality. For this reason, [29] conditioned the weight matrix of the first layer so that the diagonal blocks are equal to zeros and released this constraint after a well-chosen number of iterations.

To warranty the preservation of the unimodal information, we explore an alternative method which turned out to be better: we apply an adapted weight decay, only on the non-diagonal blocks, and decreased its contribution to the loss through time.

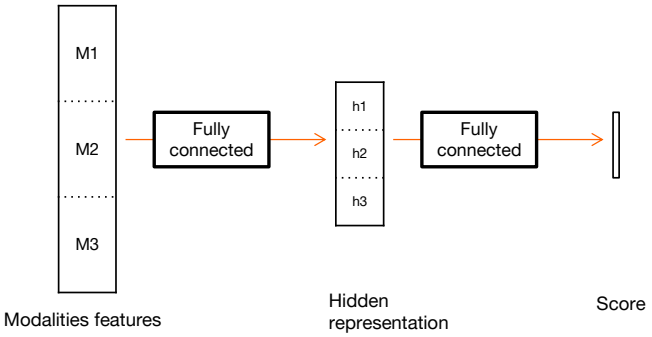

Figure 5: Fully connected model with ModDrop. The three modalities are concatenated and a fully connected with ModDrop is applied. The obtained hidden representation is then fed to a second regular fully connected layer, which will output the scores.

To be more formal, let n be the number of modalities, $W1$ be the weights matrix of the first layer. It can be divided into n^2 weights block matrices $W1_{k,l}$, modeling unimodal and intermodal contributions, with k and l, ranging over the number of modalities n. The first fully connected equation can be written as :

$$\begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix} = \begin{pmatrix} W1_{1,1} & W1_{1,2} & W1_{1,3} \\ W1_{2,1} & W1_{2,2} & W1_{2,3} \\ W1_{3,1} & W1_{3,2} & W1_{3,3} \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \\ M_3 \end{pmatrix}$$

Then, the term to add to the loss is simply (with γ_{md} decreasing through the time):

$$\gamma_{md} \sum_{k \neq l} ||W1_{k,l}||_2$$

Setting γ_{SD} to very high values during the first iterations leads to zeroing non-diagonal block matrices. Lowering it later reintroduces progressively these coefficients. From our observation, this approach provided better convergence on the considered problem.

Score trees. Our motivation is to combine the high-level information coming from the scores with the lower-level information coming from the features, all together. We did it by building what we call *Score Trees* (see Figure 6 for an illustration).

A fully connected classification neural network is applied separately to the features of the different modalities, outputting a vector of size 7. This vector is then concatenated with the scores of the two other modalities, to create a vector of size 21. A fully connected classification neural network is then fed with it and outputs a prediction vector of size 7. The aim is to make predictions with respect to the predictions coming from other modalities. Finally, these three new prediction vectors are concatenated and fed into a last fully connected classifier, which gives the overall scores. This method can be generalized to any number of modalities.

Weighted Mean. The weighted mean is the approach of the winners of the 2016 edition [18]. It consists in weighting the score of each modality and sum them up.

The weights are chosen by cross validation on the validation set, selecting the ones giving the best performance.

We applied it on the VGG-LSTM, C3D-LSTM and audio models.

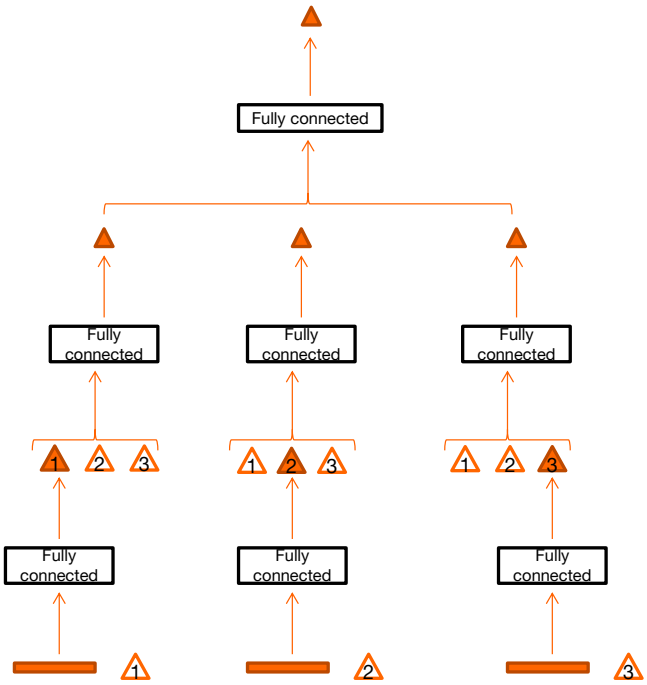

Figure 6: The *Score Tree* architecture

3 EXPERIMENTAL VALIDATION AND RESULTS

After introducing the AFEW dataset (the dataset of the challenge), this section presents the experimental validation of our method. We first present experiments done on the modalities taken separately, and then present experiments on their fusion. We finally introduce the experiments done for the EmotiW'17 challenge and the performance we obtained.

3.1 The Acted Facial Emotion in the Wild dataset

Acted Facial Emotion in the Wild (AFEW) is the dataset used by the EmotiW challenge. The 2017 version of AFEW is composed of 773 training videos, 383 validation videos and 653 test videos. Each video is labeled with one emotion among: 'angry', 'disgust', 'fear', 'happy', 'sad', 'neutral' and 'surprise'. In addition to the video, cropped and aligned faces extracted by [36, 39] are also provided.

Another important specificity of this dataset is the class distribution of the Train/Val/Test subsets (as shown in Table 1). This difference can make the performance on the Test set different from the one of the Validation set, as some classes are more challenging than others.

3.1.1 External training data. To enlarge the training set, we collected external data by selecting 380 video clips from our personal DVDs movie collection, after checking that there is no overlap between the selected movies and the ones of AFEW [12]. These movies were processed by the following pipeline: faces are first detected using our face detector (see Section 2.1 for details), and the bounding boxes and timestamps are kept. We then extract candidates temporal windows of ten seconds around all of these time stamps and asked human annotators to select and annotate the most relevant ones. To ensure the quality of our annotations, we evaluated ourselves (as human beings) on the validation set. We reached a performance from 60 % to 80 % depending on the annotator, which is compatible with the figure of 60% observed by [22].

	Training	Validation	Test
Angry	133 (17.2 %)	64 (16.7 %)	99 (15.2 %)
Disgust	74 (9.6 %)	40 (10.4 %)	40 (6.1 %)
Fear	81 (10.4 %)	46 (12 %)	70 (10.7 %)
Happy	150 (19.4 %)	63 (16.4 %)	144 (22 %)
Sad	117 (15.1 %)	61 (15.9 %)	80 (12.3 %)
Neutral	144 (18.6 %)	63 (16.4 %)	191 (29.2 %)
Surprise	74 (9.6 %)	46 (12 %)	29 (4.4 %)
Total	773	383	653

Table 1: AFEW 7.0: number of video sequences per class.

3.2 Experiments on Single Modalities

Each modality has been evaluated separately on the validation set of AFEW. The VGG-LSTM and C3D-LSTM modalities performs better than [11].

Regarding the VGG-LSTM and the C3D-LSTM, both unidirectional and bidirectional LSTM architectures [21] were evaluated, with one or two layers.

In several recent papers [3, 14], bidirectional LSTM is claimed to be more efficient than the unidirectional one and could be seen as a way to augment the data. Nevertheless, in our case, we have observed that the bidirectional LSTM was prone to over-fitting on the training set and therefore does not perform well on the validation set. The same observation has been made when increasing the number of layers. The best performing architecture is in both cases a one-layer unidirectional LSTM.

3.2.1 VGG-LSTM. The VGG model (without LSTM) has first been evaluated by taking the maximum of the scores over the sequences, giving the accuracy of 41.4% on the validation set.

Then the different LSTM architectures were tested and the best performance for each one is given in Table 2. We note a 3% improvement compared to Fan *et al.* [18]. It can be explained by the fact that our model uses the whole sequences, feeding the model with more information. Data augmentation also helps.

3.2.2 C3D-LSTM. We then experimented our method with the C3D modality alone. The performance is given in Table 3.

We observe that our implementation of C3D trained on random windows and evaluated on central windows is not as good as Fan *et al.* [18]. The C3D trained on central windows performed better but is not state-of-the-art either.

Our proposed C3D (with LSTM) has been tested with and without overlapping between windows. To evaluate the weighted C3D, the prediction of the window with the maximal softmax score among the video is first taken. It performs better without overlapping, and we observe a lower difference between training and validation accuracy. It could be explained by the fact that the number of

Method	Validation accuracy
Maximum of the scores	41.4 %
Unidirectional LSTM one layer	**48.6 %**
Bidirectional LSTM one layer	46.7 %
Unidirectional LSTM two layers	46.2 %
Bidirectional LSTM two layers	45.2 %
Fan *et al.* [18]	45.42 %

Table 2: VGG-LSTM performance on the validation set of AFEW.

Method	Validation accuracy
C3D on central window	38.7 %
C3D on random window	34 %
Weighted C3D (no overlap)	42.1 %
Weighted C3D (8 frames overlap)	40.5 %
Weighted C3D (15 frames overlap)	40.1 %
LSTM C3D (no overlap)	**43.2 %**
LSTM C3D (8 frames overlap)	41.7 %
Fan *et al.* [18]	39.7 %

Table 3: C3D and C3D-LSTM performance on the validation set of AFEW.

windows is lower if there is no overlap, the choice between the windows is therefore easier. As a second observation, the use of LSTM with the weighted C3D leads to the highest scores.

At the end, it can be observed that our C3D descriptor performs significantly better than the one of [18].

3.2.3 Audio. The audio modality gave a performance of 36.5%, lower than the state-of-the-art method (39.8%) [17]. The use of a perceptron classifier (worse than the SVM) nevertheless allowed us to use high-level audio features during fusion.

3.3 Experiments on Fusion

Table 4 summarizes the different experiments we made on fusion.

The simple baseline fusion strategy (majority vote or means of the scores) does not perform as well as the VGG-LSTM modality alone.

The proposed methods (ModDrop and Score Tree) achieved promising results on the validation set, but are not as good as the simple weighted mean on the test set. This can be explained by the largest number of parameters used for the modDrop and the score tree, and by the fact that some parameters cross validated on the validation set.

The best performance obtained on the validation set has the accuracy of 52.2 % , which is significantly higher than the performance of the baseline algorithm provided by the organizers – based on computing LBPTOP descriptor and using a SVR – giving the accuracy of 38.81 % on the validation set [10].

Fusion method	Validation accuracy	Test accuracy
Majority vote	49.3 %	_
Mean	47.8 %	_
ModDrop (sub.3)	**52.2 %**	56.66 %
Score Tree (sub.4)	50.8 %	54.36 %
Weighted mean (sub.6)	50.6 %	**57.58 %**

Table 4: Performance of the different fusion methods on the validation and test sets.

3.4 Our participation to the EmotiW'17 challenge

We submitted the method presented in this paper to the EmotiW'17 challenge. We submitted 7 runs which performance is given in Table 5).

The difference between the runs are as follows:

- Submission 2: ModDrop fusion of audio, VGG-LSTM and a LSTM-C3D with an 8-frame overlap.
- Submission 3: addition of another LSTM-C3D, with no overlap, improving the performance on the test set as well as on the validation set.
- Submission 4: fusion based on Score Trees, did not achieve a better accuracy on the test set, while observing a slight improvement on the validation set.
- Submission 5: addition of one VGG-LSTM and two other LSTM-C3D, one with 8 frames overlap and one without. These new models were selected among the best results in our random grid search on hyper-parameters according to their potential complementarity degree, evaluated by measuring the dissimilarity between their confusion matrices. The fusion method is ModDrop.
- Submission 6: weighted mean fusion of all the preceding modalities, giving a gain of 1 % on the test set, while losing one percent on the validation set, highlighting generalization issues.
- Submission 7: our best submission, which is the same as the sixth for the method but with models trained on both training and validation sets. This improves the accuracy by 1.2 %. This improvement was also observed in the former editions of the challenge. Surprisingly, adding our own data didn't bring significant improvement (gain of less than one percent on the validation set). This could be explained by the fact that our annotations are not correlated enough with the AFEW annotation.

The proposed method has been ranked 4th in the competition. We observed that, this year, the improvement of the top accuracy compared to the previous editions is small (+1.1%), while from 2015 to 2016 the improvement was of +5.2 %. This might be explained by the fact that the methods are saturating, converging towards human performance (which is assumed to be around 60 %). However, the performance of top human annotators (whose accuracy is higher than 70 %) means there is still some room for improvement.

Submission	Test Accuracy
2	55.28 %
3	56.66 %
4	54.36 %
5	56.51 %
6	57.58 %
7	58.81 %

Table 5: Performance of our submissions on the test set.

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	74,5	0	3,1	3,1	18,4	1	0
Disgust	20	**0**	5	15	40	20	0
Fear	27,1	0	34,3	1,4	21,4	15,7	0
Happy	6,9	0	0,7	82,6	6,3	3,5	0
Neutral	8,3	0	2,1	5,7	69,9	14	0
Sad	12,5	0	7,5	13,8	25	41,3	0
Surprise	21,4	0	21,4	7,1	35,7	14,3	**0**

Figure 7: Confusion matrix obtained with the seventh submission. We can see that 'disgust' and 'surprise' classes are never predicted by our model, while the three dominant classes ('happy', 'neutral' and 'angry') are well recognized. The 'neutral' class has the largest number of false positives. It underlines the difficulty, even for humans, to draw the margin between presence and absence of emotion. Rows denote true classes, columns predicted classes.

4 CONCLUSIONS

This paper proposes a multimodal approach for video emotion classification, combining VGG and C3D models as image descriptors and explores different temporal fusion architectures. Different multimodal fusion strategies have also been proposed and experimentally compared, both on the validation and on the test set of AFEW. At the EmotiW'17 challenge, the proposed method ranked 4th with the accuracy of **58.81 %**, 1.5 % under the competition winners. One important observation from this competition is the discrepancy between the performance obtained on the test set and the one on the validation set: good performance on the validation set is not a warranty to good performance on the test set. Reducing the number of parameters in our models could help to limit overfitting. Using pre-trained fusion models and, moreover, gathering a larger set of data would also be a good way to face this problem. Finally, another interesting path for future work would be to add contextual information such as scene description, voice recognition or even movie type as an extra modality.

REFERENCES

[1] Saad Ali and Mubarak Shah. 2010. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence* 32, 2 (2010), 288–303.

[2] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. 2011. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*. Springer, 29–39.

[3] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. 2012. Spatio-Temporal Convolutional Sparse Auto-Encoder for Sequence Classification. In *BMVC*.

[4] Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, and Cha Zhang. 2016. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 433–436.

[5] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. 2011. Context in emotion perception. *Current Directions in Psychological Science* 20, 5 (2011), 286–290.

[6] Lisa Feldman Barrett and James A Russell. 1999. The structure of current affect: Controversies and emerging consensus. *Current directions in psychological science* 8, 1 (1999), 10–14.

[7] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez. 2017. EmotioNet Challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210* (2017).

[8] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen. 2016. Audio visual emotion recognition with temporal alignment and perception attention. *arXiv preprint arXiv:1603.08321* (2016).

[9] JunKai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. 2014. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 508–513.

[10] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2017. From Individual to Group-level Emotion Recognition: EmotiW 5.0. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*.

[11] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2016. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 427–432.

[12] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE MultiMedia* (2012).

[13] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 423–426.

[14] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1110–1118.

[15] Paul Ekman and Wallace V Friesen. 1977. Facial action coding system. (1977).

[16] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1459–1462.

[17] Lijie Fan and Yunjie Ke. 2017. Spatiotemporal Networks for Video Emotion Recognition. *arXiv preprint arXiv:1704.00570* (2017).

[18] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 445–450.

[19] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. [n. d.]. *Learning to forget: Continual prediction with LSTM*. Technical Report.

[20] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*. Springer, 117–124.

[21] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional LSTM networks for improved phoneme classification and recognition. *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005* (2005), 753–753.

[22] Markus Kächele, Martin Schels, Sascha Meudt, Günther Palm, and Friedhelm Schwenker. 2016. Revisiting the EmotiW challenge: how wild is it really? *Journal on Multimodal User Interfaces* 10, 2 (2016), 151–162.

[23] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. 2017. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing* (2017).

[24] Pooya Khorrami, Thomas Paine, and Thomas Huang. 2015. Do deep neural networks learn facial action units when doing expression recognition?. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 19–27.

[25] Bo-Kyeong Kim, Jihyeon Roh, Suh-Yeon Dong, and Soo-Young Lee. 2016. Hierarchical committee of deep convolutional neural networks for robust facial

expression recognition. *Journal on Multimodal User Interfaces* 10, 2 (2016), 173–189.

[26] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10, Jul (2009), 1755–1758.

[27] Agata Kołakowska, Agnieszka Landowska, Mariusz Szwoch, Wioleta Szwoch, and Michał R Wróbel. 2013. Emotion recognition and its application in software engineering. In *2013 The 6th International Conference on Human System Interaction (HSI)*. IEEE, 532–539.

[28] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. 2017. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing* (2017).

[29] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. 2016. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (2016), 1692–1706.

[30] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. 2015. Deep Face Recognition.. In *BMVC*, Vol. 1. 6.

[31] Robert Plutchik and Henry Kellerman. 2013. *Theories of emotion*. Vol. 1. Academic Press.

[32] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos.. In *NIPS*. 568–576.

[33] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[34] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Dennis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 3–10.

[35] Peter Washington, Catalin Voss, Nick Haber, Serena Tanaka, Jena Daniels, Carl Feinstein, Terry Winograd, and Dennis Wall. 2016. A wearable social interaction aid for children with autism. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2348–2354.

[36] Xuehan Xiong and Fernando De la Torre. 2013. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 532–539.

[37] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. 2016. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *IEEE Transactions on Affective Computing* (2016).

[38] Anbang Yao, Junchao Shao, Ningning Ma, and Yurong Chen. 2015. Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 451–458.

[39] Xiangxin Zhu and Deva Ramanan. 2012. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2879–2886.