Metric Learning Based Multimodal Audio-visual Emotion Recognition

Esam Ghaleb, Mirela Popa and Stylianos Asteriadis

Abstract—People express their emotions through multiple channels such as visual and audio ones. Consequently, automatic emotion recognition can be significantly benefited by multimodal learning. Even-though each modality exhibits unique characteristics, multimodal learning takes advantage of the complementary information of diverse modalities when measuring the same instance, resulting in enhanced understanding of emotions. Yet, their dependencies and relations are not fully exploited in audio-video emotion recognition. Furthermore, learning an effective metric through multimodality is a crucial goal for many applications in machine learning. Therefore, in this paper, we propose Multimodal Emotion Recognition Metric Learning (MERML), learned jointly to obtain a discriminative score and a robust representation in a latent-space for both modalities. The learned metric is efficiently used through the Radial Basis Function (RBF) based Support Vector Machine (SVM) kernel. The evaluation of our framework shows a significant performance, improving the state-of-the-art results on the eNTERFACE and CREMA-D datasets.

Index Terms—Multimodal learning, Metric learning, Audio-video emotion recognition, Convolutional neural networks, Fisher vectors

1 Introduction

During the last years, thanks to the recent advances in sensing technologies, multimodal learning has attracted a notable body of research in the field of machine learning [1]. For both humans and machines, information obtained from diverse sensory modalities, when observing any psycho-social experiences, gives compelling results in understanding these phenomena. To that end, there is a need to develop systems which benefit from the nature of the targeted task and the complementary data provided by various modalities. However, this is not a trivial task, since different modalities have their unique data distributions and distinct statistical properties [1]. Therefore, a multimodal paradigm needs to be developed, to capture the informative aspects of the present channels and build an efficient representation taking into consideration possible non-linearities in their correlations, as well. Applications of multimodal learning include, but are not limited to, person identification, emotion recognition [2], and multimedia retrieval [1].

In emotion recognition, due to the complex and varied expressions of emotions across individuals, and the high dimensionality of audio-video data, the role of multimodal learning becomes even more evident. The goal here is to predict high-level affective content from low-level human-centered signals such as video and audio. The usage of different modalities can help in understanding those highly sophisticated sub-conscious experiences, making emotion recognition one of the most interesting areas of multimodal learning [2]. Studies show that human perception of intended emotions increases when both audio and video modalities are available [3]. For example, emotions such as fear, disgust, and surprise require simultaneous audio-visual cues for higher recognition rates. In addition, audio can be more crucial for understanding anger, while facial expressions play a bigger role in happiness recognition [3].

In the literature, most of the recent research studies on multimodal emotion recognition are limited either to feature concatenation of various representations or late fusion of individual modalities in combination with various classification algorithms [2]. However, by employing an efficient metric distance, the accuracy of many classification and retrieval problems can be increased [4], as this can contribute to obtaining improved performance and more robust representation. Metric learning approaches learn distances to bring similar inputs closer and dissimilar ones further apart, which are more discriminative than the conventional Euclidean distance. This transformation is done through convex optimization with pairwise constraints [4]. Typical examples include Large Margin Nearest Neighbour (LMNN) [4], Geometric Mean Metric Learning (GMML) [5] and Information Theoretical Metric Learning (ITML) [6].

Motivated by the success of unimodal metric learning and the multimodality nature of emotion recognition, we propose a novel Multimodal Emotion Recognition Metric Learning (MERML) framework, which leverages the audio-visual information to jointly learn modality-specific Mahalanobis metrics. The approach simultaneously optimizes the distance and similarity metrics for audio and video representations, such that they are more discriminative in the learned latent-space, contributing to better emotion classification.

A substantial benefit of our proposed learning approach relies on its generalization ability. In other words, it can be applied in various multimodal learning domains beyond emotion recognition, for jointly learning and fusing complementary data channels. The contributions of the paper are summarized as follows:

- We propose MERML for the challenging audio-video emotion recognition task (Section 3). This methodology achieves state-of-the-art results in two benchmarks: Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [3] and eNTERFACE

- *E. Ghaleb, M. Popa and S. Asteriadis are with Department of Data Science and Knowledge Engineering, Faculty of Science and Engineering, Maastricht University. Address: Bouillonstraat 8-10, 6211 LH Maastricht, The Netherlands.*
  E-mail: {esam.ghaleb, mirela.popa, stelios.asteriadis}@maastrichtuniversity.nl.

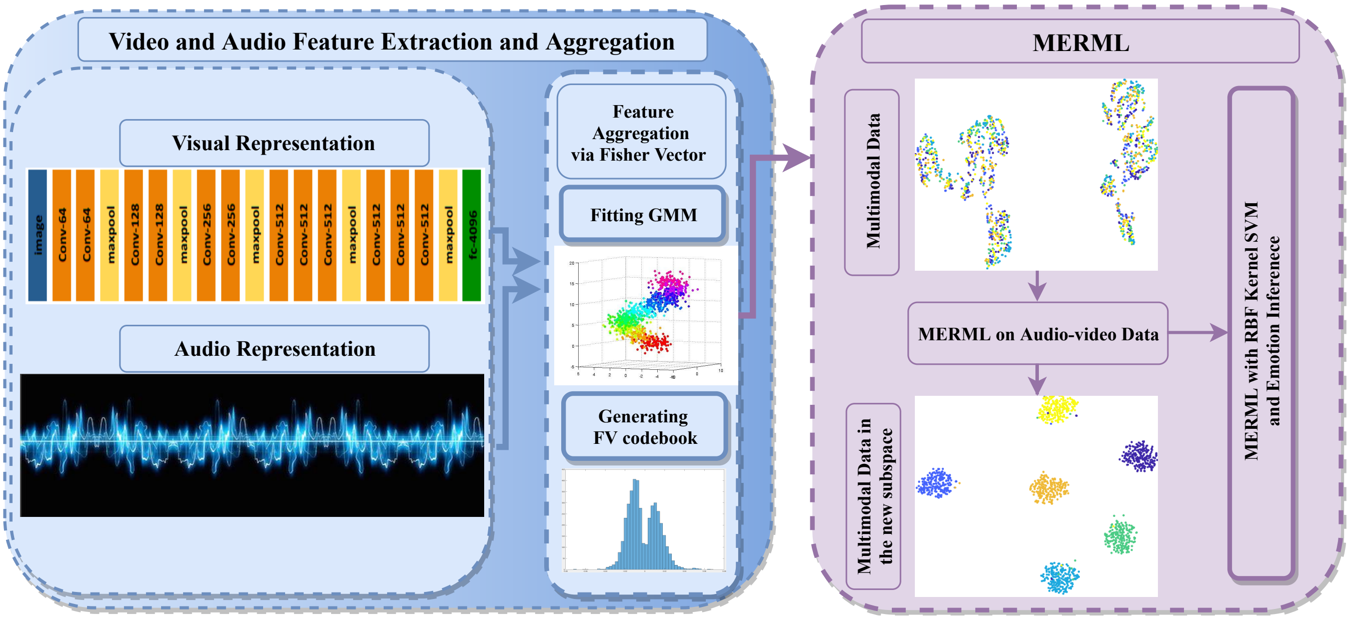


Fig. 1: The proposed method starts with extracting visual and audio features and aggregating them via Fisher vectors. Then, MERML implementation is applied, with subsequent incorporation through SVM for emotion recognition.

[7]. We jointly learn the modality-specific metric, that aims to not only capture and explain the complex relationship between these two modalities, but also to efficiently learn a latent-space for improved representations and enhanced classification.

- The proposed distance is incorporated efficiently in RBF-based SVM, benefiting the emotion classification task. Besides, the developed MERML is scalable, since it learns modality-specific metrics without imposing any constraint on them, such as the dimension of their features or the number of data samples. Furthermore, the rationale of the proposed distance is intuitive, contributing to the explainability of the model, in terms of modalities' input importance for each emotion. These details are further elaborated in Section 3.5.
- The efficacy of the proposed method is evaluated extensively on two public benchmarks for audio-video emotion recognition, CREMA-D [3], and eNTERFACE [7]. We provide extensive experimental results, showing a notable performance of our method in comparison with baseline metric learning approaches such as LMNN [4], GMML [5], ITML [6] or other methods (Section 5).

The proposed methodology is depicted in Figure 1, where first, we extract audio-visual features and then aggregate them through Fisher vectors to obtain spatio-temporal representations (Section 4). Subsequently, the MERML is learned, followed by the video emotional content inference phase, discussed in Section 5.

2 Related Work

Audio-video Emotion Recognition: During the past decades, many researchers investigated human emotional states through different channels and descriptions [2]. In the literature, the description of the emotional content can be roughly divided into two subcategories: discrete and continuous models. Maybe the most typical representative discrete model is that of the Ekmanian emotions, which describes the six basic, universal emotions of sadness, anger, happiness, surprise, disgust, and fear. On the other side, typical dimensional models include the valence and arousal space. Usually, datasets annotated with emotions are annotated according to a particular model, for example, CREMA-D [3] and eNTERFACE [7], which are annotated on the discrete, Ekmanian emotions, and RECOLA database [8] based on a continuous model. Many of previous studies, such as those in [9], [10], focus on concatenating the audio-visual representations or on the late fusion of individual modalities, while our approach efficiently learns a distance metric by capturing non-linearities. Furthermore, in this work, we propose a joint multimodal metric learning for audio-visual emotion recognition. We efficiently learn this distance metric by capturing the non-linearity between the high-dimensional modalities, contributing to obtaining an enhanced performance.

Moreover, with the recent improvements in Deep Learning (DL) [2], deep architectures have become popular and useful in extracting high-level features from multimodal data, especially for facial and speech emotion recognition [10]. Most of the works that use deep learning for video-audio emotion recognition employ separate architectures for each data channel.

In our work, we adopt the deep learning representation through CNN features for the visual modality [11]. Generally speaking, despite the great performance of deep learning in many areas of pattern recognition, affective computing still lacks the potential of taking the full advantages of this technology, due to the inadequacy of properly annotated data and the limited ability of existing architectures to explain the cause-effect relationships in such data[1]. Like other traditional automated human affect recognition works, our study benefits from the well understood analytical models, such as distance metric learning and Support Vector Machines (SVM).

Metric Learning (ML): Many studies have implemented ML in a unimodal context for a plethora of domains, such as multimedia retrieval, computer vision, and machine learning [4], [6]. In [4], authors defined LMNN metric, such that k-*nearest*

[1] https://sites.google.com/site/dalcom2017cvpr/

neighbors of a sample always belong to the same class. They effectively obtain a Mahalanobis distance metric as the solution to a semi-definite program. GMML [5] learns the distance metric using a smooth unconstrained convex optimization formulation. In [6], authors proposed ITML algorithm to minimize the differential relative entropy between two multivariate Gaussians conditioned with a constraint in the distance function. In addition, authors in [12], proposed Heterogeneous Domain Adaptation (HDA) to augment audio modality with visual features. Their purpose is to investigate the knowledge transfer between audio and visual domains using modality augmentation via ML.

Likewise, multimodal metric learning has shown success in miscellaneous domains utilizing various representations of visual data [13], as well as other data channels such as text. A simple way to learn a metric for multimodal data is by concatenating the features of the modalities and then applying classical metric learning such as LMNN, GMML or ITML.

3 Proposed Method

3.1 Definitions and Notions

Audio-Visual Emotion Recognition (AVER): Given a dataset $\mathbb{D}$ with audio-visual emotional content, consisting of m samples, each annotated with a discrete emotion y_i:

$$\mathbb{D} = \{(x_1^v, x_1^a, y_1), (x_2^v, x_2^a, y_2), \ldots, (x_m^v, x_m^a, y_m)\}$$

where $x_i^v \in X^{d^v}$ and $x_i^a \in X^{d^a}$ denotes video and audio feature vectors corresponding to the i^{th} sample of video $X^{d^v \times m}$ and audio $X^{d^a \times m}$ data matrices, while $y_i, i \in \{1, \ldots, N_e\}$ refers to the given discrete emotion label. The goal, in AVER, is to predict the emotional content of a given sample test. In this paper, we use uppercase letters to denote a matrix, e.g. $X \in \mathbb{R}^{m \times d}$, and lowercase letters for vectors, e.g. $x \in \mathbb{R}^d$. Additionally, we define the following operators and terms:

- m: the number of samples.
- d^a, d^v: the dimensionality of audio and video features.
- $M^a = W^{a^T} W^a$ and $M^v = W^{v^T} W^v$: $M^a$and M^v are distance matrices for audio and video modalities, while W^a and W^v are the linear transformation matrices .
- p^v and p^a: the dimensionality of the new subspace.
- y_{ij}: is 1 if i and j belong to the same class, or -1 otherwise.
- S and D: two sets of similar (positive) and dissimilar (negative) instance pairs:

$$S = \{(x_i, x_j, y_{ij}) | y_{ij} = 1\}, \text{ and } D = \{(x_i, x_j, y_{ij}) | y_{ij} = -1\}$$

- $d(x_i, x_j)$: the multimodal distance of two samples x_i and x_j given their audio and video modalities.

3.2 A Brief Review of Metric Learning (ML)

Standard distance metric assumes a higher similarity between two samples' representations of a similar class and bigger difference otherwise. Given two samples, x_i and x_j, the standard Euclidean distance is:

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T (x_i - x_j)} \tag{1}$$

In standard distance ML, the goal is to find an optimal metric M according to the similarity and dissimilarity constraints. This is done through a convex optimization, to learn the transformation matrix W for the original features, such that $M = W^T W$, where M is symmetric and positive. As a result, the new formulation of equation 1 is:

$$d_M^2(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j) = (x_i - x_j)^T W^T W (x_i - x_j) \tag{2}$$

In more details, ML modifies the standard Euclidean distance to improve its discriminative ability, such that the distance between similar classes would be as small as possible, while enlarging it otherwise. Another benefit of ML includes dimensionality reduction of the feature vectors, by the linear projection $W \in \mathbb{R}^{p \times d}$, where $p \ll d$, and $p \geq rank(W)$. Note that, if $M = I^{d \times d}$, where $I^{d \times d}$ denotes the identity matrix, then the metric is reduced to a Euclidean distance.

3.3 Multimodal Emotion Recognition Metric Learning (MERML) Formulation

In this section, we define MERML for audio-visual emotion recognition or similar tasks of a multimodal nature. Like the conventional unimodal ML, similar samples should be projected, in the new space, as close as possible to each other, while dissimilar ones must be placed further apart from each other. We aim to jointly learn modality-specific metrics, M^a and M^v, that maximize the prediction accuracy utilizing both audio and video data channels:

$$\begin{aligned} d_{W^v, W^a}^2(x_i^v, x_j^v, x_i^a, x_j^a) &= d_{W^v}^2(x_i^v, x_j^v) + d_{W^a}^2(x_i^a, x_j^a) \\ &= \|W^v x_i^v - W^v x_j^v\|_2^2 + \|W^a x_i^a - W^a x_j^a\|_2^2 \end{aligned} \tag{3}$$

Both the dimensionality of modalities' feature vectors, $d^v$and d^a, and the linear transformation matrices, W^v and W^a, in the new subspace can be different. The aim is to learn Mahalanobis matrices $M^v = W^{v^T} W^v$ and $M^a = W^{a^T} W^a$, using a convex formulation in the low-rank subspace, such that $W^v \in R^{p^v \times d^v}$and $W^a \in \mathbb{R}^{p^a \times d^a}$. These two matrices can also serve in reducing the high dimensionality of audio-visual modalities $X^v \in R^{d^v}$ and $X^a \in R^{d^a}$, which makes the developed methodology applicable for large-scale datasets.

In addition, a non-negative weighting scheme is applied in MERML. This strategy is specifically useful in audio-visual emotion recognition, as the varied contribution of audio-visual modalities in emotion prediction has been supported in many studies. For example, in [3], visual information has been reported to have more impact on the final decision. Furthermore, the assigned weighting scheme, $(\omega,\ 1 - \omega)$ can further help the algorithm to converge faster:

$$d_{W^v, W^a}^2(x_i^v, x_j^v, x_i^a, x_j^a) = \omega d_{W^v}^2(x_i^v, x_j^v) + (1 - \omega) d_{W^a}^2(x_i^a, x_j^a) \tag{4}$$

In these new subspaces, the weighted distance of both modalities and the linear projections are more efficient, such that using both audio and face, the distance between two samples i and j is smaller than a learned threshold $b \in R$, if they belong to the same class or larger otherwise. As stated in [4], this condition can be further imposed with a margin larger than 1 by the following constraint:

$$y_{ij}\left(b - d_{W^v, W^a}^2(x_i^v, x_j^v, x_i^a, x_j^a)\right) > 1 \tag{5}$$

3.4 Optimization

To solve the defined formulation of MERML in equation 4 with the weighting scheme and the constraint condition, the following hinge-loss function is applied and optimized through Stochastic Gradient Descent (SGD):

$$\begin{aligned}\underset{W^v,W^a,\omega,b}{\operatorname{argmin}} L_{ij} = \sum_{i,j} max\Big[1 - y_{ij}\Big(b - \big[\omega(x_i^v - x_j^v)^T W^{v^T} W^v (x_i^v - x_j^v) \\ + (1-\omega)(x_i^a - x_j^a)^T W^{a^T} W^a (x_i^a - x_j^a)\big]\Big), 0\Big]\end{aligned} \tag{6}$$

This loss-function is non-differentiable due to the max-operation. However, the sub-gradient is usually used instead, through the following condition:

$$\begin{aligned}\nVdash = (y_{ij}\Big(b - [\omega(x_i^v - x_j^v)^T W^{v^T} W^v (x_i^v - x_j^v) \\ + (1-\omega)(x_i^a - x_j^a)^T W^{a^T} W^a (x_i^a - x_j^a)]\Big) > 0)\end{aligned} \tag{7}$$

Where $\nVdash$ is a logical operator, and it is *0* if the condition holds or *1* otherwise. The sub-gradient is applied to solve for W^v, W^a, b, and ω. Therefore, through an on-line SGD, at each iteration t, based on pairs of audio-visual samples (i, j), either negative or positive with the same frequency.

3.5 Classification: MERML in RBF Kernel SVM

Following the metric learning computation, we apply the learned distance, through the kernel trick in SVM. To that end, the standard Euclidean distance function $d^2(x_i, x_j)$ in the RBF kernel (equation 8) [14] is replaced by the proposed multimodal metric distance:

$$K(x_i, x_j) = \exp\left(-\frac{d^2(x_i, x_j)}{2\sigma^2}\right) \tag{8}$$

Consequently, the classification is carried out by SVM with the new distance. Therefore, in our proposal of MERML, we efficiently employ the multimodal distance to obtain the following non-negative kernel function:

$$K(x_i^v, x_j^v, x_i^a, x_j^a) = \exp\left(-\frac{d^2_{W^v,W^a}(x_i^v, x_j^v, x_i^a, x_j^a)}{2\sigma^2}\right) \tag{9}$$

The motivations behind replacing RBF kernel by the proposed metric can be listed as follows:

- Concatenating the projected features of each modality and then using SVM affects the potential discriminative power of the learned metric and results in an expensive computation. In our experiments, the performance was less accurate, than with the proposed metric. In other words, instead of using a classification algorithm on the concatenated features resulted from the linear projections of different modalities, we utilize the weighted score in the kernel trick directly.
- This approach makes the method scalable. By scalability, we refer to the properties of the introduced method, which projects the data in a latent space, having a finite dimension, enabling a compact representation and being able to deal with big sizes of training data. Plugging the scores, resulted from each modality in equation 9, allows an efficient emotion inference. Specifically, in AVER, for some instances, the face might not be detected, or vice-versa the audio may be missing. For example, the audio modality is highly probably to be missing in instances associated with emotions related to happiness, which is manifested mainly through facial expressions.

  We evaluated this setting in a heterogeneous inference. In particular, we compared the performance of the two modalities before and after MERML. For example, we have noticed that, there is an improvement on the visual modality performance compared to its performance prior to MERML. While, we noticed that, audio modality has slightly better or similar results than those prior to MERML. This improvement is due to the fact that the approach manages to capture the complementary elements of the multimodal data during the training phase.

3.6 Positive and Negative Sample Mining in MERML

One of the main challenges in ML is that the optimization through SGD often suffers from slow convergence when applied on a large scale dataset. Therefore, we propose an optimized process of selecting positive-negative pairs, using several criteria. When the emotional classes are not balanced, we ensure that an equal percentage of samples from all emotional classes are always present during training. Then, using the confusion matrix between emotions, we balance the ratio of difficult vs. easy pairs. We consider difficult pairs, the ones belonging to emotions which are difficult to be discriminated, and they are especially relevant for MERML for obtaining an efficient high-level representation. Furthermore, the sampling is done using the multimodal information, such that the MERML distance defined in equation 4 selects hard negative samples (x_j^v, x_j^a) from both modalities, that maximize the $d^2_{W^v,W^a}(x_i^v, x_j^v, x_i^a, x_j^a)$ distance.

4 Feature Extraction and Aggregation

Face Tracking and Alignment: For a given video sample, face and facial landmark tracking is done using the Ensemble of Regression Trees (ERT) method described in [15]. ERT provides robust and accurate landmark positions in challenging conditions, such as varying illumination and poses, with reliable and robust tracking in real-time. In a face track, each face is aligned by registering it with respect to facial landmarks, for instance, eye's centers, nose, mouth, and chin, to a canonical frame through a similarity transformation. Facial images are cropped and re-sized to a fixed resolution: 224 x 224.

Visual Features: we use a CNN representation based on the VGG-face model [11]. VGG-face is a 16-layer CNN model, trained with 2.6M facial images of 2.6K people for face recognition in the wild. As the model is trained for the facial recognition task, and not on emotional data, we fine-tune the model, by keeping the convolutional and the Fully Connected Layer (FCL) layer 6, discard the remaining FCLs and append 1000 and 7 dimensional layers with softmax classification and multinomial logistic loss. For fine-tuning, we use the training set of the Facial Emotion Recognition dataset (FER) [16], and for validation, we use FER public test set. In the feature extraction stage, we employ the FCL 6's output as the facial signature. This layer outputs a 4096-d feature vector.

Audio Features: we extract audio features through utilizing the speech analysis toolkit openSMILE [17]. This open-source library is popular and widely used for extracting audio signals that capture both voice quality and prosodic characteristics of the speaker. We extract a set of features as explained in [18], including the following descriptors: 34 energy & spectral related Low-Level Descriptors (LLDs)×21 functionals, 4 voicing related LLDs×19 functionals, 34 delta coefficients of energy & spectral LLDs×21 functionals, 4 delta coefficients of the voicing related LLDs×19 functionals and 2 voiced/unvoiced durational features. The functionals computed on both energy/spectral and voicing related LLDs include: arithmetic mean, standard deviation, skewness, kurtosis, quartiles, quartile ranges, percentile 1%, 99%, percentile range, position max/min, up-level time 75/90, linear regression

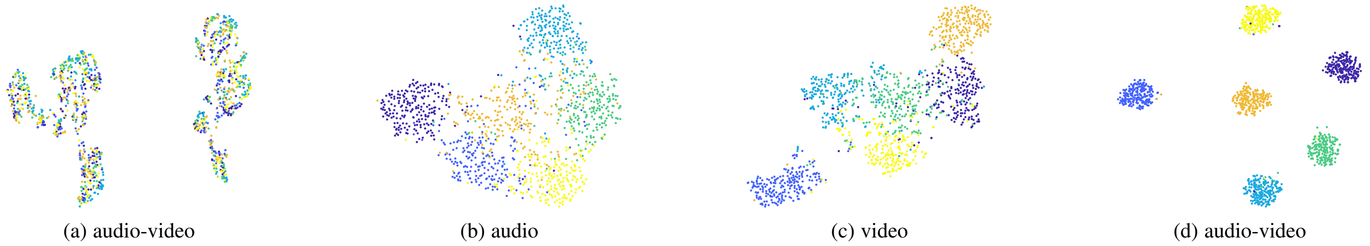


Fig. 2: t-SNE embedding of the eNTERFACE features, in the original subspace and in the new learned subspace. Features in (a) are concatenated from both modalities in the original space. Note that they are highly correlated and the number of formed clusters is not well defined. In (b), (c) and (d), we visualize the audio, video and audio-video data following MERML, where the clusters are better structured. The cluster colors represent the emotional classes. (The figure is better viewed in color.)

coefficients, and linear regression error (quadratic/absolute). For each video-clip, the resulted feature vector dimension is 1582.

Temporal Feature Aggregation using Fisher Vectors: as the extracted set of features has high dimensions, (e.g. each video frame has a 4096-d feature vector), and each modality description has its unique characteristics and data distribution, we employ a high-level representation, through Fisher Vector (FV) encoding. FV is used for aggregating and clustering low-level features of each frame (e.g. CNN and audio features), obtaining one FV from all the frames in a sequence (e.g. video-clip's face track), by fitting a parametric generative model such as GMM to the features. GMM can be referred to as a *probabilistic visual vocabulary*, while FV encodes the gradient of the local descriptors log-likelihood with respect to the GMM parameters. More details about Fisher Vector Encoding can be found in [19].

FV encoding is used to obtain a compact and single feature vector for each modality of a given video-clip. For example, in our work, for a given AVER dataset ($\mathbb{D}$), the audio and visual features are extracted and then aggregated via FV for each sample (i). Therefore, x_i^v and x_i^a denote video and audio FVs corresponding to the i^{th} sample of video $X^{d^v \times m}$ and audio $X^{d^a \times m}$ data matrices.

5 EXPERIMENTAL RESULTS

To illustrate the efficacy of the proposed method, we present an extensive experimental evaluation and report the results on two benchmarks of audio-video emotion recognition datasets: CREMA-D [3] and eNTERFACE [7]. As each database has a different number of emotion categories and due to their varying settings and recording setups, we provide an independent evaluation for each dataset.

In the experiments on each dataset, we present the following results:

1) Unimodal evaluation of each representation separately using RBF-SVM, showing the initial performances of the employed audio and video features in a unimodal emotion recognition task.
2) The results of the baseline metric learning techniques LMNN [4], ITML [6] and GMML [5] on the concatenated representations of audio-video features, followed by RBF-SVM on the representations obtained from LMNN.
   In this scenario, the concatenated video-audio features are the same as the ones in RBF-SVM baseline, and they are later on used separately in MERML. In this way, we provide a fair comparison for MERML against these two baselines.
3) The results of MERML on the pairs of audio-video (AV) features according to its formulation in Section 3, illustrating the benefits of the proposed method.

First, we qualitatively evaluate the proposed MERML by visualizing the projected data in the newly learned subspace. For this purpose, we utilize t-SNE, a popular visualization, and unsupervised dimensionality reduction tool [20]. Figure 2 shows eNTERFACE sample features before and following MERML. We can observe that prior to MERML, the emotional classes are highly correlated, hence form similar clusters. However, applying MERML improved the cluster forming based on the emotional content, leading to well-separated emotions in the new subspace.

5.1 Sensitivity Analysis

Since the performance of MERML depends on W_a and W_b projection matrices, as formulated in Equation 4, we conduct a sensitivity analysis with regards to these variables. In particular, we check how the initialization of W_a and W_b and their dimensions contribute to the overall performance. We tested two initialization approaches, random and PCA initialization with various dimensions. Fig 3 details the sensitivity analysis on both datasets. We notice that, in the case of CREMA-D, PCA initialization of W_a and W_b gives better performance. However, various dimensions of these projection matrices give very close results, especially, when these dimensions are in a range of 150 to 250. This shows that MERML is robust under different configurations. In the rest of our experiments using MERML, we use PCA initialization while the dimensions of W_a and W_b are set to 200.

5.2 Evaluation on CREMA-D

CREMA-D [3] is a multimodal emotion expression dataset and contains 7442 clips from 91 actors (43 females and 48 males). Their age ranges between 20 and 74 and they come from a variety of races and ethnicities (African American, Asian, Caucasian, Hispanic). Actors were asked to speak 12 sentences, according to six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) with four different levels (Low, Medium, High and Unspecified).

The binomial majority recognition for video only, audio-only and audio-visual presentations, which are 40.9%, 58.2%, and 63.6% respectively. In addition, the traditional majority recognition, where more than 50% of the raters selected a specific emotion, for audio-only, video-only, and audio-video modalities are 45.5%, 69.0%, and 74.8% respectively.

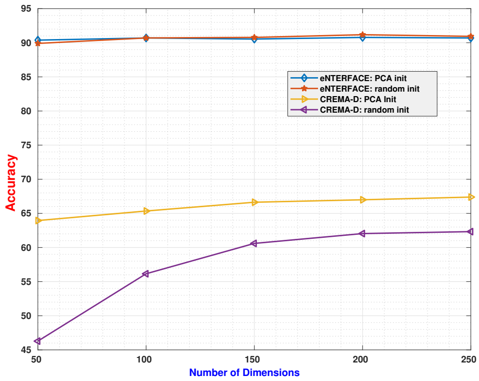


Fig. 3: Sensitivity analysis on W_a and W_v with regards to their initialization and dimensions.

5.2.1 Results on CREMA-D

To the best of our knowledge, as there is not a technical evaluation baseline for this dataset, we divided the dataset into 10-folds to perform cross-validation based on subjects (the actors). In other words, for each fold, subjects' clips are either in the training or testing sets, and there is not any overlap between these two sets. Subsequently, in each fold, we train the whole pipeline of FV encoding, learning MERML and SVM on the training folds, and then test it on the remaining fold. The reported results are the average of these 10-folds.

Positive and negative mining: Due to the size of CREMA-D, the number of possible positive and negative pairs is huge, approximately 45 Million pairs. For this reason, we followed a careful selection of the used pairs during the SGD iterations for MERML optimization. Firstly, we put more emphasis on the positive samples of the same emotion, e.g. anger, to be selected from different subjects. Secondly, we pooled more negative pairs of emotions from the same subject. In this way, we ensured that the MERML training focus is on capturing the intra-class and inter-class variations depending on the emotions, rather than the subjects of the video clips.

5.2.2 Uni-modal and multimodal interaction:

For a better understanding of the contribution of the audio and video modalities, Fig. 4 (a) provides the uni-modal and MERML results of the employed features, in terms of average accuracy and the accuracy on each emotion of CREMA-D. The first bar in each group gives the average result of MERML on the audio-visual representations. MERML was able to capture the complementary and supplementary information of both modalities, achieving an accuracy of 65.5%. In addition, MERML outperformed the individual modalities and yielded in a performance increase of 9.3% and 16.7% over the audio-only and video-only perception, respectively. In addition, as shown in figure 4 (a), the recognition accuracy of each emotion (with a slight exception for fear) increased by the multimodal learning through MERML. These results show that while MERML on audio-visual is the leading modality, the contribution of its sub-modalities varies in some emotions. For example, the audio is more significant than the facial expressions for anger, and video has more impact in recognizing happiness. On the other hand, disgust and neutral emotions require both data channels for better recognition, proving the benefits of multimodal learning in any case.

In terms of unimodal results, audio outperforms visual representations on CREMA-D. This result is due to the limited number of sentences in CREMA-D, which led to learning good prosodic features and vocal expressions for each emotion. In addition, in [3], authors reported that only for the audio modality, emotion perception increased by 10% when raters interacted and responded to more clips, while the interaction did not have an impact for the video modality.

5.2.3 Multimodal evaluation:

MERML results and the comparisons between its performance against LMNN, ITML, GMML, and SVM-RBF baselines are given in TABLE 1, showing that MERML resulted in at least 1.3% performance gain. In addition, MERML approach outperformed both human-perception based on binomial-majority vote and the recently published results in [21]. In [21], the performance (65.0% accuracy) was obtained by combining facial and audio temporal features with Long-Short-Term-Memory (LSTM). These results show the efficiency of our approach for an enhanced joint multimodal learning and fusion.

5.3 Evaluation on eNTERFACE

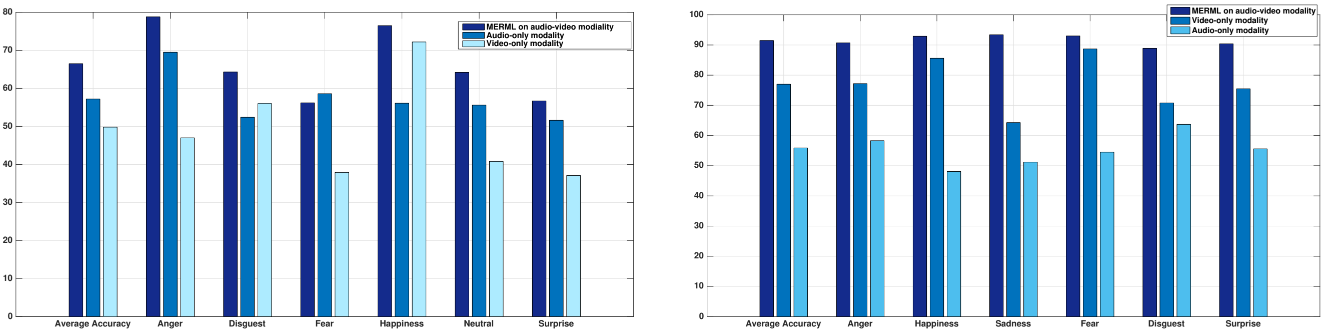
eNTERFACE is a multimodal dataset which contains six archetypal emotions: anger, happiness, disgust, fear, surprise, and sadness. It includes 42 subjects, who were asked to simulate the emotions in 5 different reactions, resulting in 1260 video recordings. 23% of the recordings are obtained from women and 77% are from men, including respondents of diverse cultural backgrounds. In the evaluation, we follow the protocol in [9] by splitting each class members into a balanced 10 folds for cross-validation. Consecutively, in each fold, we train the whole pipeline of FV encoding, learning MERML and SVM on the training folds, and then testing it on the remaining fold.

5.3.1 Uni-modal and multimodal interaction:

Fig. 4 (b) provides the detailed performance of the visual, audio and MERML modalities for each emotion of eNTERFACE. On average, MERML obtains an accuracy of 91.5%, which helped to increase the performance of individual modalities by 14.5% and 35.6% for visual-only and audio-only perceptions, respectively. Furthermore, the fusion of audio-visual modalities through MERML boosted the performance by improving the classification accuracy for each emotion. In particular, performance for sadness was largely better following MERML. This shows how the presence of both modalities is important in multimodal perception and subsequently emotion recognition.

5.3.2 Multimodal evaluation:

TABLE 1 presents the average recognition accuracies for the eNTERFACE dataset. We compare MERML with the current state-of-the-art audio-visual approaches in [9], [10] on eNTERFACE. Our MERML approach on audio-visual representations, gives the highest recognition accuracy (91.5%), which is competitive to the state-of-the-art result (89.4%) reported by [10], that was obtained through end-to-end DL methods, namely, 3D CNN (C3D) cascaded with deep-belief networks (DBN). The reason behind this improvement is that DL methods have more parameters to train and learn, which needs to be properly initialized, e.g. via transfer learning.



(a) CREMA-D

(b) eNTERFACE

Fig. 4: Bar diagrams for average and per-emotion performance on CREMA-D and eNTERFACE, to show the multimodal and unimodal accuracies of RBF-SVM on audio-only, RBF-SVM on video-only, and MERML on audio-video. For example, the first three bars indicate the average accuracy of the three modalities in each dataset.

CREMA-D		eNTERFACE	
Methods and features	**Accuracy %**	**Methods and features**	**Accuracy %**
Dual Attention with LSTM: AV [21]	65.0	Score-level bimodal SVM [9]	87.4
Human Perception (binomial majority recognition)	63.6	Late fusion on C3D-DBN [10]	89.4
SVM on the concatenated AV representations	65.2	SVM on the concatenated AV representations	84.7
ITML on the concatenated AV representations	60.5	ITML on the concatenated AV representations	77.5
GMML on the concatenated AV representations	65.0	GMML on the concatenated AV representations	81.9
LMNN on the concatenated AV representations	63.5	LMNN on the concatenated AV representations	85.1
MERML (Section 3) on AV	**66.5**	MERML (Section 3) on AV	**91.5**

TABLE 1: The average recognition accuracy of MERML and other methods on CREMA-D and eNTERFACE.

Interestingly, we can observe that MERML increases the performance of the audio-visual representations, compared with the case when using only SVM or LMNN, GMML, ITML on the concatenation of audio and video representations. To illustrate this fact, the recognition rate of audio-visual representations increased by 6.8%, 6.4%, 9.6% and 14% when using MERML in comparison with SVM, LMNN, GMML, and ITML respectively. In addition, comparing the baseline of LMNN, GMML, ITML or SVM and MERML using significance testing, we were able to validate the substantial gains (p-value $\ll 0.05$). These results show that MERML can provide a distance measure that enhances the performance of audio-video emotion classification.

5.4 Multimodal Interactions:

For a better explanation of MERML and the contribution of the audio and video modalities, we check the agreement in emotion prediction, based on unimodal and multimodal representations obtained via MERML. Our analysis shows that while the audio-visual is the leading modality, the contribution of its sub-modalities varies in each emotion. There are three categories in terms of modalities' contribution. The first category contains anger and sadness emotions, where audio is more significant than the facial expressions. In this case, audio contributes to more than 20% of the multimodal recognition. The second category includes happiness and disgust, where video has more impact on recognizing these emotions, which could be at least 10% better than audio. On the other hand, in the third category, fear and neutral require both data channels for better recognition.

6 Conclusions and Future Work

We presented a joint multimodal metric learning for audio-video emotion recognition. MERML can be applied in various multimodal contexts in which data complementarity could be exploited for increasing the performance, through an improved latent-space data representation. Our approach exploited successfully the dependencies and the complementary information of audio and video modalities in the context of emotion recognition, as their representations are well structured in the newly learned subspace, and their mutual emotion recognition is maximized. The quantitative and qualitative evaluation of the method on two datasets, utilizing distinct pairs of visual and audio representations, demonstrated the significant contribution of the method to an increased classification accuracy, achieving better than the state-of-the-art results in eNTERFACE and CREMA-D datasets. Furthermore, the comparison with the LMNN, GMML and ITML baseline metric learning approaches showed the benefits of our method, which is efficiently learning the two modalities and optimizes their contribution for an enhanced performance.

Acknowledgment

This work has been funded through H2020-MaTHiSiS project under Grant Agreement No. 687772. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Nvidia TITAN XP GPUs used for this research.

References

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[2] P. V. Rouast, M. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing*, 2019.

[3] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[4] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.

[5] P. H. Zadeh, R. Hosseini, and S. Sra, "Geometric mean metric learning," in *Proc. of the 33rd Int. Conf. on Machine Learning (ICML)*, 2016, pp. 19–24.

[6] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. of the 24th Int. Conf. on Machine learning*, 2007, pp. 209–216.

[7] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proc. 22nd Int. Conf. on*. IEEE, 2006, pp. 8–8.

[8] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE Int. Conf. on*. IEEE, 2013, pp. 1–8.

[9] G. Chetty, W. Michael, and G. Roland, "A multilevel fusion approach for audiovisual emotion recognition," *Emotion Recognition: A Pattern Analysis Approach*, pp. 437–460, 2015.

[10] D. Nguyen, K. Nguyen, S. Sridharan *et al.*, "Deep spatio-temporal features for multimodal emotion recognition," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, 2017, pp. 1215–1223.

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition." in *BMVC*, 2015.

[12] C. Athanasiadis, E. Hortal, and S. Asteriadis, "Bridging face and sound modalities through domain adaptation metric learning," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2019)*, 2019.

[13] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan, "Large margin multi-metric learning for face and kinship verification in the wild," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 252–267.

[14] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[15] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 1867–1874.

[16] I. J. Goodfellow, D. Erhan, P. L. Carrier *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Int. Conf. on Neural Information Processing*. Springer, 2013, pp. 117–124.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. of the 18th ACM Int. Conf. on Multimedia*. ACM, 2010, pp. 1459–1462.

[18] M. Valstar, J. Gratch, B. Schuller *et al.*, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. of the 6th Int. Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[19] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *Int. Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.

[20] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[21] R. Beard, R. Das, R. W. Ng, P. K. Gopalakrishnan, L. Eerens, P. Swietojanski, and O. Miksik, "Multi-modal sequence fusion via recursive attention for emotion recognition," in *Proc. of the 22nd Conf. on Computational Natural Language Learning*, 2018, pp. 251–259.