High-Level Geometry-based Features of Video Modality for Emotion Prediction

Raphaël Weber
CentraleSupelec/FAST/IETR
Avenue de la Boulaie
35576 Cesson-Sévigné,
France
raphael.weber@
centralesupelec.fr

Vincent Barrielle
Dynamixyz
80, avenue des Buttes de
Coesmes
35700 Rennes, France
vincent.barrielle@
dynamixyz.com

Catherine Soladié
CentraleSupelec/FAST/IETR
Avenue de la Boulaie
35576 Cesson-Sévigné,
France
catherine.soladie@
centralesupelec.fr

Renaud Séguier
CentraleSupelec/FAST/IETR
Avenue de la Boulaie
35576 Cesson-Sévigné,
France
renaud.seguier@
centralesupelec.fr

ABSTRACT

The automatic analysis of emotion remains a challenging task in unconstrained experimental conditions. In this paper, we present our contribution to the 6th Audio/Visual Emotion Challenge (AVEC 2016), which aims at predicting the continuous emotional dimensions of arousal and valence. First, we propose to improve the performance of the multimodal prediction with low-level features by adding high-level geometry-based features, namely head pose and expression signature. The head pose is estimated by fitting a reference 3D mesh to the 2D facial landmarks. The expression signature is the projection of the facial landmarks in an unsupervised person-specific model. Second, we propose to fuse the unimodal predictions trained on each training subject before performing the multimodal fusion. The results show that our high-level features improve the performance of the multimodal prediction of arousal and that the subjects fusion works well in unimodal prediction but generalizes poorly in multimodal prediction, particularly on valence.

1. INTRODUCTION

Automatic emotion analysis has gained a constantly growing interest during the last decades. Indeed, applications can be seen in different domains such as healthcare or human-computer interface.

Two types of problems are usually considered: emotion recognition among discrete categories or emotion prediction into continuous dimensions. Both problems try to take advantage of signal processing methods to automatically analyse emotions in accordance with emotional representation models developed by psychologists. On one hand, classification methods are used for emotion recognition - generally the discrete categories are the prototypical expressions defined by Ekman [9]. On the other hand, regression methods are used to infer the emotional state, which is defined as a vector of continuous emotional dimensions such as valence, arousal, power or expectancy. This dimensional representation first appeared in psychological research [28] and has been later studied with principal component analysis [11].

Many modalities have been studied to solve those problems. The most popular ones are audio, video and physiological signals. The reader can refer to [18] for a review of the existing features extraction for emotion recognition with those modalities and to [38] for a survey on both audio and visual modalities and their fusion.

Since non-verbal cues are an important part of communication [23] and affect display, great interest has been carried to facial expression analysis [25]. Excellent performance is now achieved with in-the-lab data, but the challenge remains opened for in-the-wild data [21]. Indeed, in-the-wild data implies additional noise that cannot be tackled by algorithms trained on laboratory-controlled data. When it comes to facial expression analysis, it includes large head pose variation or occlusions.

Thus, research is now focusing effort on emotion analysis in real-world conditions [21]. The 6th Audio/Visual Emotion Challenge and Workshop (AVEC 2016) addresses this problem. One of the sub-challenges is to predict emotional dimensions with multimodal natural data.

In this paper, we propose a high-level geometry-based features extraction of the video modality carrying information about head pose and facial expression. The head pose is estimated by fitting a 3D reference mesh to the 2D facial landmarks. The facial expression features are extracted from the 2D facial landmarks projected on a person-specific model [32]. The novelty of our work is the unsupervised computa-

AVEC'16, October 16 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2988257.2988262

tion of the model thanks to an automatic neutral face detection. The originality of this person-specific model lies in its self-organization of a continuous invariant representation of facial expressions [32]. Thus, even if the model is person-specific, the information contained in models of different persons can be compared. The originality of our approach is to map a data-driven representation into the arousal/valence representation driven by psychologists' models.

Another contribution is the fusion of unimodal regressions along training subjects: by training a regressor for each training subject, we can weight their respective contribution on the prediction by fusing them.

The remaining of this paper is organized as follow: Section 2 introduces the related work, Section 3 details our method for high-level features extraction and subjects fusion, Section 4 gives the experimental results in the framework of the AVEC 2016 challenge and Section 5 concludes the paper.

2. RELATED WORK

As head pose and facial expression give relevant information on the emotional state, we briefly review in this section methods for head pose estimation and facial features extraction for facial expression analysis.

2.1 Head pose estimation

In the following, we will describe a few methods for head pose estimation that has been recently used. For an exhaustive survey, the reader can refer to [24].

Head pose estimation gives high-level information on the head gesture or the person activity. So, it is adapted to the prediction of an emotional dimension such as arousal that reports the state of being awake. In [13] the head gestures are used as features to map into 5 emotional dimensions, including arousal and valence. We distinguish three approaches for head pose estimation: manifold learning, flexible model and geometric approaches.

In the manifold learning approach, the goal is to map the facial features into a low-dimensional continuous manifold. Different dimensionality reduction methods such as principal component analysis (PCA) [22, 30] or linear discriminative analysis (LDA) [5, 34] have been used to learn the manifold. The weakness of these approaches is the assumption that head pose is mainly responsible for dimensionality reduction, which is not guaranteed.

The flexible model approach estimates the pose by fitting a deformable model to the facial image [19, 16, 1]. The parameters of the model are trained on a set of data with different head pose configuration. The performance of this approach is directly linked to the generalization ability of the deformable model and the accuracy of the facial features used for the fitting.

In the geometric approach, head pose is estimated by minimizing the distances between the facial landmarks and a projected 3D model [15, 37, 2]. The accuracy of the head pose estimation depends on the robustness of the facial landmarks localization among factors such as image resolution.

2.2 Facial expression features

Among the popular cues to infer an emotional state, facial expression has been widely explored since the 1990s as it is shown in the survey [25]. A more recent survey can be found in [21]. The facial features extraction is a crucial step in facial expression classification or regression. In the following, we will review some of the recent facial expression features extraction methods.

Since facial expression results from the displacements of the facial muscles, the geometry-based facial features are popular for facial expression analysis. Indeed, the localization and tracking of facial landmarks give direct information about the facial muscles activity. Various methods have been developed for the tracking of 2D facial landmarks such as active appearance models (AAM) [6], particle filtering [26] or supervised descent method (SDM) [36]. For facial expression analysis, these low-level features need landmarks alignment to get rid of the head pose for instance, but this is still a challenging task for large head pose variation. To tackle this problem, features based on angles and distances between facial landmarks have been proposed [17, 7, 39, 14].

The manifold-based representation learns a high-level representation of facial features by learning how to map low-level features to a manifold. Many methods emerged to perform this learning, among them local linear embedding (LLE) [4], locality preserving projections (LPP) [29], modified Lipschitz embedding [3] or multiple manifold learning [35]. The weakness of this approach is that it is highly dependent on the distribution of the data. If a test expression has not been learned during the training phase, the representation of this expression fails.

More recently, deep learning received a growing interest for facial expression features extraction. The idea is to extract high-level facial features with a deep convolutional neural network (CNN) and use the resulting features for facial expression analysis. In [20] the CNN is fed with both 3D geometric facial features and 2D appearance facial features, whereas in [8] the CNN is directly fed with pixel images of faces. So far, those methods gave promising results for facial expression analysis in unconstrained experimental conditions.

3. METHOD FOR EMOTIONAL DIMENSION PREDICTION

3.1 Overview of the system

As shown in the figure 1, our system reproduces the AVEC 2016 baseline architecture [33]. For several modalities we extract features and then learn how to map independently those features into the emotional dimensions of arousal and valence with a regression task. Once the unimodal regressors are trained, we learn how to fuse the unimodal predictions with another regression task.

The contributions are:

- The computation of high-level geometry-based features for the video modality in addition to the features proposed in the baseline: head pose and facial expression features, that we call "expression signature".

  To compute the high-level geometry-based features, we use 2D facial landmarks extracted from the video thanks to the Supervised Descent Method [36]. This method tracks 49 2D facial landmarks by minimizing an error between the pixel image and a deformable generic shape model. Those geometric features give information about the morphology, the expression and the head pose. Our high-level features extraction allows extracting the head pose and the expression separately while removing the morphology.

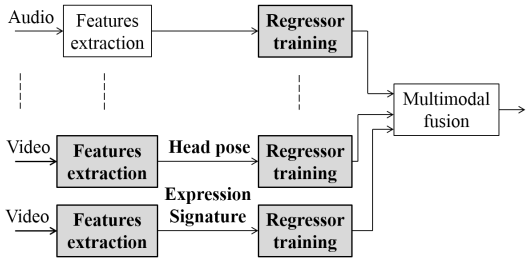

Figure 1: Overview of the baseline system [33] for the prediction of one emotional dimension. Our contributions take action in the gray boxes. In addition to the baseline we compute high-level geometry-based features (head pose and expression signature). We also adopt a fusion scheme on the unimodal regressor training.

On one hand, the head pose is estimated by fitting a reference 3D face mesh to the 2D facial landmarks. On the other hand, we compute an unsupervised person-specific model into which we project the facial 2D landmarks. We obtain a 3D vector containing the expressive information that we call "expression signature".

- A strategy of late fusion on the training subjects for unimodal regressors to predict the emotional dimensions - arousal and valence.

  We train one regressor for each training subject and each modality and then fuse the training subjects for each modality with a regression on both train set and development set.

3.2 High-level geometry-based features extraction

The figure 2 gives an overview of the high-level geometry-based features extraction used for the video modality. Beforehand we estimate the head pose from the 2D landmarks, this defines the first part of our high-level features. For the facial expression, we compute the unsupervised person-specific model of the facial deformations. The only information needed from the subject is the neutral face, so we detect it automatically and once it is done, the model can be computed. The facial expression is then defined as the projection of the 2D landmarks in the model. It gives a 3D vector that we call "expression signature" containing the expressive information.

Afterwards, for both head pose and expression signature, we concatenate the high-level features with their derivatives using central difference and several sample distances. Thus we combine static and dynamic features. Then we use the same process as in the AVEC 2016 baseline [33] by computing for each frame the mean and the standard deviation of the features on a centred temporal window and standardise the resulting features.

3.2.1 Head pose estimation

The head pose estimation is done by fitting the pose of a reference 3D mesh to the 2D facial landmarks. We note x_0 the reference 3D mesh and l the 2D landmarks from which the pose is estimated.

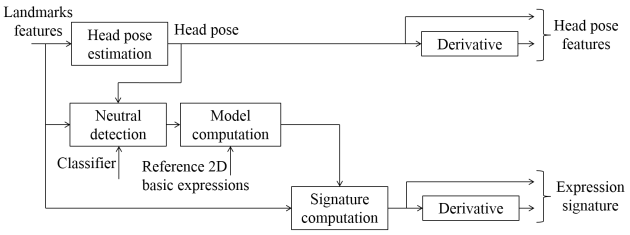

Figure 2: Overview of the geometry-based high-level features extraction: head pose features and expression signature.

Thus, we find scale s, rotation matrix R and translation t of the head pose by minimizing:

$$E_{pose}(s, R, t) = ||sPR(x_0 + t) - l||^2 \quad (1)$$

where: $P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ is the orthographic projection matrix in the X-Y plane. The optimisation is computed using the Gauss-Newton algorithm.

We then define the head pose features as the concatenation of s, the 3 angles of R and the translation of t on the X and Y axes.

3.2.2 Automatic neutral face detection

The person-specific model, needed for the second part of our high-level features extraction, represents the facial deformations around the neutral face. So we have to detect automatically the neutral face in order to compute the model in an unsupervised manner.

Let X_i, X_j and X_k be three facial landmarks. For two points X_u and X_v we note $Y_{u,v} = X_u - X_v$. The angle is computed as the inverse tangent of the cosine over the sine of the angle between $Y_{i,j}$ and $Y_{k,j}$.

Let X be the n facial landmarks of dimension p, we note $X_{i,j}$ the j-th coordinate of the i-th landmark and $\bar{X}$ the mean landmark of X. To compute the distance, we first compute the scale s of the face defined as the mean distance between the landmarks and the mean landmarks:

$$s = mean_{i \in (1,n)} \left(\sqrt{\sum_{j=1}^{p} (X_{i,j} - \bar{X})^2} \right) \quad (2)$$

Thus we get the normalized landmarks $\tilde{X}$ and we compute the distance between the landmarks X_l and X_m as:

$$d_{l,m} = \sqrt{\sum_{j=1}^{p} (\tilde{X}_{l,j} - \tilde{X}_{m,j})^2} \quad (3)$$

The design of our angle-distance features consists of defining the triplets (X_i, X_j, X_k) for the angle computation and the pairs (X_l, X_m) for the distance computation. We chose those triplets and pairs by features-engineering with respect to their ability to discriminate facial expressions. Our angle-distance features are composed of 3 angles and 6 distances on the eyebrows region, 4 angles on the eyes region, 8 angles and 3 distances on the mouth region.

The figure 3 illustrates our neutral face detection scheme. We train 4 classifiers with angle-distance features, each one

for a specific facial region: the whole face with 6 classes (neutral, anger, disgust, joy, sadness, surprise), the eyebrows region with 3 classes (neutral, raised, frowned), the eyes region with 2 classes (opened, closed) and the mouth region with 6 classes (neutral, anger, disgust, joy, sadness, surprise).

We choose 6 classes for the whole face and the mouth region because it allows discriminating a neutral face from an expressive face in the most important facial deformation directions.

The decisions from the 4 classifiers are then fused: a neutral face is detected if all classifiers output the neutral class. So, our approach is hybrid in the sense that it combines both global information (for instance a prototypic expression) and local information (for instance action units). In this case, the fusion is designed to detect a neutral face but it could be applied to detect any prototypic expression. Furthermore, the head pose is taken into account so that a non-frontal face is rejected.

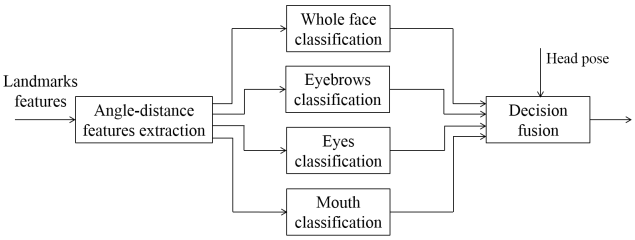

Figure 3: Overview of the automatic prototypic expression detection. We fuse global (whole face) and local (eyebrows, eyes and mouth) classifiers to detect the prototypic expression. We use it for neutral face detection.

3.2.3 Signature computation in the unsupervised person-specific model

The high-level facial expression features are computed by projecting the 2D facial landmarks in an unsupervised person-specific model. The advantage of the model is to provide a continuous invariant representation of facial expressions [32]. Thus expressions of different subjects projected in their respective model can be compared. Moreover, having a person-specific allows a more accurate analysis than a generic model [31].

It is based on the assumption that facial deformations are organized the same way for everyone with the neutral face as the central expression. Let us assume that a set of different subjects performs a neutral face and several facial expressions and that for each expression the 2D facial landmarks are recorded and then aligned. For each subject we perform a principal component analysis (PCA) on the expressions aligned features. If we observe the projections of the expressions in the PCA space, the expressions are organized the same way. The figure 4 shows such an example. Two subjects performed a neutral face and the same 8 expressions.

The main idea of the model is to take profit of the PCA ability to extract main deformations of data. We need to define the expressions to perform so that each one spans specific facial deformations with the neutral face being in the centre of the deformations.

The figure 5 illustrates the different steps of the model

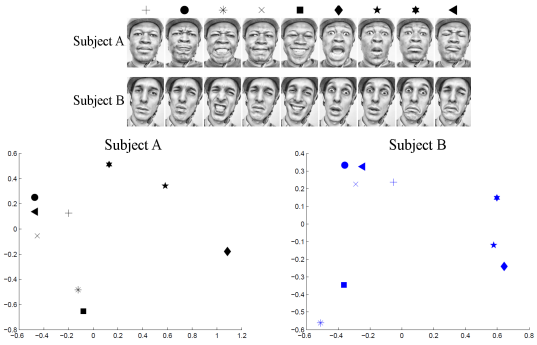

Figure 4: Example of expressions self-organization after a principal component analysis on landmarks features for two subjects, the expressions landmarks are projected in the PCA space and the first 2 axes are shown.

computation. The neutral is automatically detected with the method introduced previously. Then we create 5 plausible expressions landmarks of the subject with piece-wise affine warping [12]: the deformations from neutral face to the 5 expressions of a reference subject are used to warp the automatically detected neutral face [32]. We call the obtained expressions "basic expressions".

PCA is computed on the landmarks of the neutral face and the basic expressions after having aligned them. When projected in the PCA space, the basic expressions are always organized the same way around the neutral. We thus obtain a continuous space spanning the deformation directions contained in the basic expressions.

To take profit of this self-organisation, we perform a Delaunay tessellation on the projected neutral and basic expressions. This gives an invariant structure of simplexes where the neutral is connected to every basic expression. Since this structure is invariant, we can map the PCA space into a normalized space, called "signature space", where the neutral is at the origin and the basic expressions lie in a specific location on a sphere surface.

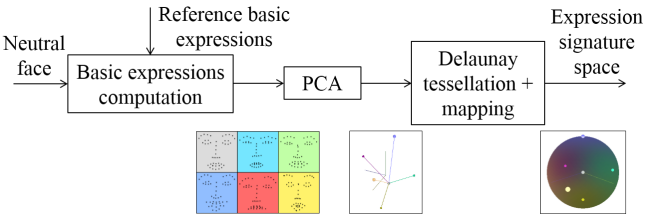

Figure 5: Unsupervised person-specific model computation. A principal component analysis is computed on the neutral face and the basic expressions landmarks. Then a Delaunay Tessellation is performed on the PCA vectors of the neutral face and the basic expressions in order to map the PCA space into an expression signature space with an invariant structure.

We choose the five basic expressions given in the table 1 for the stability that they confer to the structure: anger, disgust, joy, sadness and surprise. For each basic expressions,

we give the local deformation in the eyebrows region, the eyes region and the mouth region.

Table 1: Basic expressions for the computation of the person-specific model. The local deformations of the eyebrows, the eyes and the mouth are given.

Basic expression	Eyebrows	Eyes	Mouth
Anger	Frowned	Opened	Shrinked
Disgust	Frowned	Squinting	Horizontally stretched
Joy	Neutral	Tightening	Corners raised
Sadness	Raised	Closed	Corners lowered
Surprise	Raised	Wide opened	Opened

The figure 6 illustrates the invariant structure of the signature space. Two subjects perform a smiling expression which is projected in their respective person-specific model. In the central part of the figure 6, the smile projected in the PCA space of each subject is displayed by a black point. We can see it is located closely to the basic expression of joy in each subject space. Then this space is mapped into the signature sphere and thanks to the invariant structure of the model, the projected smile, still displayed as a black point, is now located in the same zone of the sphere. Thus we can analyse the expressions of different subjects in an unsupervised manner.

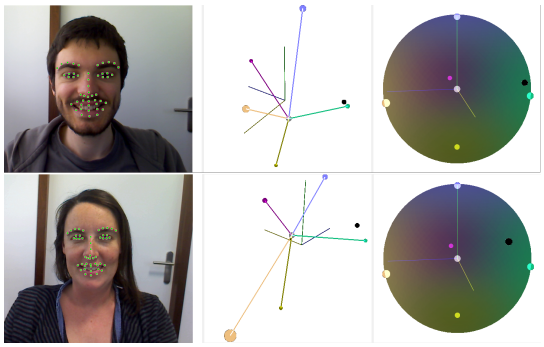

Figure 6: Illustration of the invariant expression signature space - the smile of 2 subjects is projected in their respective PCA space (black point), when mapped into the expression signature space the projected smile is located in the same zone of the sphere for both subjects.

3.3 Full system integration and fusion

The goal of our global system is to show that multimodal prediction from low-level features benefits from the addition of our high-level geometry-based features as new modalities.

On one hand, the expression signature space is a data-driven unsupervised 3D space. On the other hand, the emotional dimensions arousal and valence is a supervised space based on psychological modelling. With a regression method, we learn how to map the expression signature space into the arousal/valence space. We claim that our signature space is coherent with the arousal/valence representation and thus the expression signature features are relevant to predict those emotional dimensions.

To illustrate this point, we computed the expression signature of the subjects of the RECOLA train set and development set [27] and displayed them in the figure 7 with a color shade going from yellow to blue corresponding to an increasing value of valence. Since our signature space structure is invariant, there is no need to align the expression signatures between the subjects. We can see that the neutral expression signature located at the origin gather most of the valence equal to zero (green points), whereas positive valence (blue points) tends to go in a direction corresponding to the basic expression of joy. On the other hand, most of the negative valence (yellow points) is concentrated in the opposite direction of the basic expression of joy.

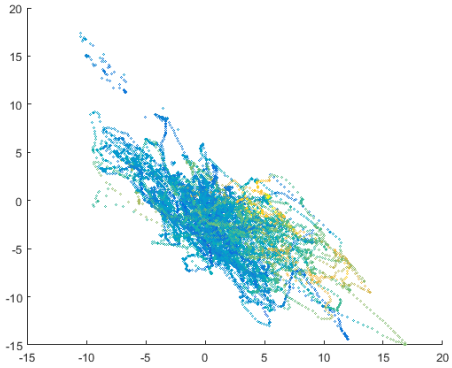

Figure 7: First two axes of the expression signatures of the subjects of the RECOLA train set and development set [27]. The colour shade goes from yellow to blue for low valence to high valence. It shows that the expression signature space gives a good representation of emotion.

We adopt a strategy of late fusion over the training subjects for unimodal regression, that we call "subjects fusion". Thus we can balance the contribution of each training subject in the unimodal prediction of the emotional dimensions.

For each training subject, we train a unimodal regressor the same way as in the baseline system [33] and we predict the emotional dimension for both the train set and the development set. Then we train a regressor with the predictions obtained with both training and development subjects in order to map them into the emotional dimension.

4. EXPERIMENTAL RESULTS

4.1 Experimental setup

The goal of our global system is to predict the emotional dimensions of arousal and valence with our high-level geometry-based features in addition to the features provided by the AVEC 2016 baseline [33].

The dataset that is used for the AVEC 2016 challenge is the RECOLA database [27]. There is a set of 27 subjects that is equally split in 3 three subsets: train set, development set and test set.

The AVEC 2016 baseline [33] has been reproduced in the programming language Python. We used the linear Support

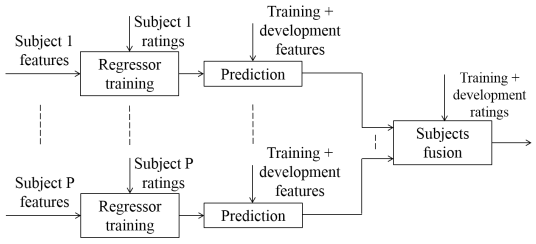

Figure 8: Overview of the fusion over the training subjects for one modality. It is done for each modality and each emotional dimension. A regressor is trained for each of the P training subjects, then the predictions of the emotional dimension are computed for both the train set and the development set and fused together with a regression method.

Vector Machine (SVM) of the liblinear library [10] for the regression training at every step of our method. Concerning the unimodal training we ignored the frames with no features. The complexity of the SVM was optimised in the range $[10^{-5}-10^{4}]$ for the subjects fusion and the multimodal fusion.

In addition to the multimodal features provided, we computed our high-level geometry-based features and used them for emotion prediction. We concatenated the features with their derivatives for the sample distances 1, 5 and 10. As in the AVEC 2016 baseline [33], we computed the mean and the standard deviation of our high-level features on a centred temporal window and then standardise them. The size of the window and the choice of standardisation technique have been optimised on the development set.

The automatic neutral detection has been processed for every subject of the dataset in order to compute their person-specific model and then compute the expression signatures. Among all the neutral faces detected in the video of each subject, we chose the one with the minimal head pose.

4.2 Performance

We first compared the performances of the unimodal emotion prediction with the baseline geometric features and our high-level geometry-based features. Then we studied the contribution of our high-level geometry-based features on the performance of the multimodal emotion prediction.

In a second time, we analysed the performance of our subjects fusion on the unimodal and multimodal emotion prediction.

4.2.1 High-level vs. low-level geometry-based features

On the development set, we computed the performance of the AVEC 2016 baseline system [33] for the low-level geometric features provided by [33] and our high-level geometry-based features, namely head pose and expression signature. The table 2 shows the results. The performance is defined as the concordance correlation coefficient (CCC, see [33]).

The results of the unimodal prediction are reported in the first part of the table. On arousal, the best performance is achieved with the expression signature features. We can also see that the performances with head pose features and geometric features are comparable. On valence, the head pose features perform poorly, and the best performance is achieved with the geometric features, while the performance with the expression signature features is quite close.

It is interesting to note that the geometric features provided in the AVEC 2016 baseline [33] and our expression signature features perform similarly, whereas their dimension is 632 and 24 respectively.

The second part of the table 2 gives the results of multimodal prediction. On the development set, we conducted three schemes of multimodal fusion. Firstly, we fused the predictions of all the modalities provided in the AVEC 2016 baseline [33] (marked as "Baseline" in the table 2). Secondly, we removed the geometric features from the baseline modalities and we added our high-level geometry-based features in the multimodal fusion (marked as "Baseline minus geometric + high-level" in the table 2). Finally, we fused the predictions from all the AVEC 2016 baseline modalities [33] and from our high-level geometry-based features (marked as "Baseline + high-level" in the table 2).

The results on the development set show that the low-level geometric features and our high-level geometry-based features perform similarly in the multimodal fusion. We also give the results of the third fusion scheme on the test set.

Table 2: Concordance correlation coefficient (CCC) of the unimodal and multimodal prediction of arousal and valence on the development and test set. An asterisk points out when our high-level features are used. The modality "geometric" corresponds to the geometric features provided in [33]. For multimodal fusion, we use the term "baseline" for all the features provided in [33] and "high-level" for our head pose and expression signature features. The baseline has been reproduced in programming language Python, so the results on the baseline slightly differ from those provided in [33].

Dataset	Modality(ies)	Arousal	Valence
Development	Geometric	0.397	**0.603**
Development	Head pose*	0.354	0.228
Development	Expression signature*	**0.471**	0.562
Development	Baseline	0.786	**0.672**
Development	Baseline minus geometric + high-level*	**0.800**	0.625
Development	Baseline + high-level*	0.791	0.662
Test	Baseline + high-level*	0.663	0.645

4.2.2 Subjects fusion

To assess the efficiency of our subjects fusion, we computed the performance of the unimodal prediction with and without the subjects fusion on the development set. The table 3 shows the results, the performance is still defined as the concordance correlation coefficient (CCC). We can see that for both arousal and valence dimensions, the subjects fusion improves the performance for almost all the modalities.

Then, we used the unimodal predictions with subjects fusion in order to fuse all the modalities of the AVEC 2016 baseline [33] and our high-level features. We conducted the experiment on the development and the test set. Moreover, we adopted two schemes of multimodal fusion: on one hand, we fused the unimodal predictions on the train set, and on

Table 3: Concordance correlation coefficient (CCC) of the unimodal prediction of arousal and valence on the development set. We compare the baseline features and our high-level features with and without subjects fusion. An asterisk points out when our high-level features are used. The subjects fusion improves the results for most of the modalities on both arousal and valence.

	Arousal		Valence	
Features	Without subjects fusion	With subjects fusion	Without subjects fusion	With subjects fusion
Audio	0.793	**0.818**	0.450	**0.456**
ECG	0.299	**0.468**	0.160	**0.221**
HRHRV	0.381	**0.424**	0.319	**0.413**
EDA	0.099	**0.187**	0.201	**0.281**
SCL	0.112	**0.197**	0.123	**0.277**
SCR	0.084	**0.193**	0.103	**0.174**
Appearance	0.508	**0.594**	0.493	**0.506**
Geometric	0.397	**0.476**	0.603	**0.683**
Head pose*	0.354	**0.434**	0.228	**0.298**
Expression signature*	0.471	**0.520**	**0.562**	0.531

the other hand, we fused the unimodal predictions on both train set and development set.

The table 4 shows the results. As expected, the fusion of both the train set and the development set improves the performance on the development set. However, it generalizes badly on the test set since the performance slightly decreases on the test set compared to the multimodal fusion on the train set alone.

With subjects fusion, the performance of the multimodal fusion is slightly better on arousal than without subjects fusion, but it performs poorly on valence. It shows that the use of the development set in the training does not give good generalization on the test set.

Table 4: Concordance correlation coefficient (CCC) of the multimodal prediction of arousal and valence with subjects fusion on the development and test set. The multimodal fusion is either trained with the predictions on the train set or on the concatenation of the train set and development set. We used all the modalities provided in [33] and our high-level features.

Testing set	Learning set	Arousal	Valence
Development	Training	0.857	0.700
	Training + Development	**0.861**	**0.719**
Test	Training	**0.682**	**0.468**
	Training + Development	0.681	0.448

5. CONCLUSION

In this work, we proposed high-level geometric based features for the prediction of the emotional dimensions of arousal and valence with the video modality. In the framework of the AVEC 2016 baseline [33], the results showed that our high-level features perform similarly to the low-level geometric features in multimodal prediction, and even slightly better on arousal.

Moreover, we presented a strategy of late fusion over the training subjects that allows balancing the contribution of each training subject in the unimodal prediction of the dimensional emotion. The results showed that this approach works well for most of the modalities on the development. However, it generalizes badly on the test set for multimodal fusion.

Future work could focus on improving the subjects fusion for unimodal prediction or on a features selection scheme for multimodal fusion so that redundant information is ignored.

6. ACKNOWLEDGMENTS

This work was supported by the FUI MILES.

7. REFERENCES

[1] T. Baltrušaitis, P. Robinson, and L.-P. Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2610–2617. IEEE, 2012.

[2] C. Canton-Ferrer, J. R. Casas, and M. Pardàs. Head pose detection based on fusion of multiple viewpoint information. In *International Evaluation Workshop on Classification of Events, Activities and Relationships*, pages 305–310. Springer, 2006.

[3] Y. Chang, C. Hu, R. Feris, and M. Turk. Manifold based analysis of facial expression. *Image and Vision Computing*, 24(6):605–614, 2006.

[4] Y. Chang, C. Hu, and M. Turk. Manifold of facial expression. In *AMFG*, pages 28–35, 2003.

[5] L. Chen, L. Zhang, Y. Hu, M. Li, and H. Zhang. Head pose estimation using fisher manifold learning. In *AMFG*, pages 203–207, 2003.

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001.

[7] A. Dapogny, K. Bailly, and S. Dubuisson. Pairwise conditional random forests for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3783–3791, 2015.

[8] W. Deng, J. Hu, S. Zhang, and J. Guo. Deepemo: Real-world facial expression analysis via deep learning. In *2015 Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2015.

[9] P. Ekman and H. Oster. Facial expressions of emotion. *Annual review of psychology*, 30(1):527–554, 1979.

[10] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[11] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.

[12] A. Goshtasby. Piecewise linear mapping functions for image registration. *Pattern Recognition*, 19(6):459–466, 1986.

[13] H. Gunes and M. Pantic. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *International conference on intelligent virtual agents*, pages 371–377. Springer, 2010.

[14] D. Han, N. Al Jawad, and H. Du. Facial expression identification using 3d geometric features from microsoft kinect device. In *SPIE Commercial+ Scientific Sensing and Imaging*, pages 986903–986903. International Society for Optics and Photonics, 2016.

[15] T. Horprasert, Y. Yacoob, and L. S. Davis. Computing 3d head orientation from a monocular image sequence. In *25th Annual AIPR Workshop on Emerging Applications of Computer Vision*, pages 244–252. International Society for Optics and Photonics, 1997.

[16] C. Hu, J. Xiao, I. Matthews, S. Baker, J. F. Cohn, and T. Kanade. Fitting a single active appearance model simultaneously to multiple images. In *BMVC*, pages 1–10, 2004.

[17] M. Huang, G. Ngai, K. Hua, S. Chan, and H. V. Leong. Identifying user-specific facial affects from spontaneous expressions with minimal annotation. 2014.

[18] A. Konar, A. Halder, and A. Chakraborty. Introduction to emotion recognition. In *Emotion Recognition: A Pattern Analysis Approach*. John Wiley & Sons, Inc., 2014.

[19] N. Krüger, M. Pötzsch, and C. von der Malsburg. Determination of face position and pose with a learned representation based on labelled graphs. *Image and vision computing*, 15(8):665–673, 1997.

[20] H. Li, J. Sun, D. Wang, Z. Xu, and L. Chen. Deep representation of facial geometric and photometric attributes for automatic 3d facial expression recognition. *arXiv preprint arXiv:1511.03015*, 2015.

[21] B. Martinez and M. Valstar. Advances, challenges, and opportunities in automatic facial expression recognition. In B. S. M. Kawulok, E. Celebi, editor, *Advances in Face Detection and Facial Image Analysis*, pages 63 – 100. Springer, 2016.

[22] S. J. McKenna and S. Gong. Real-time face pose estimation. *Real-Time Imaging*, 4(5):333–347, 1998.

[23] A. Mehrabian. Communication without words. *Psychological today*, 2:53–55, 1968.

[24] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626, 2009.

[25] M. Pantic and L. J. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1424–1445, 2000.

[26] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 97–102. IEEE, 2004.

[27] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *Proceedings of EmoSPACE 2013, held in conjunction with FG 2013*, Shanghai, China, April 2013. IEEE.

[28] J. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[29] C. Shan, S. Gong, and P. W. McOwan. Appearance manifold of facial expression. In *International Workshop on Human-Computer Interaction*, pages 221–230. Springer, 2005.

[30] J. Sherrah, S. Gong, and E.-J. Ong. Understanding pose discrimination in similarity space. In *BMVC*, pages 1–10. Citeseer, 1999.

[31] C. Soladié, N. Stoiber, and R. Seguier. A new invariant representation of facial expressions: Definition and application to blended expression recognition. In *2012 19th IEEE International Conference on Image Processing*, pages 2617–2620. IEEE, 2012.

[32] C. Soladié, N. Stoiber, and R. Séguier. Invariant representation of facial expressions for blended expression recognition on unknown subjects. *Computer Vision and Image Understanding*, 117(11):1598–1609, 2013.

[33] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge. In *Proceedings of AVEC'16, co-located with ACM MM 2016*, Amsterdam, The Netherlands, October 2016. ACM.

[34] J. Wu and M. M. Trivedi. A two-stage head pose estimation framework and evaluation. *Pattern Recognition*, 41(3):1138–1158, 2008.

[35] R. Xiao, Q. Zhao, D. Zhang, and P. Shi. Facial expression recognition on multiple manifolds. *Pattern Recognition*, 44(1):107–116, 2011.

[36] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013.

[37] Y. Xiong and F. Quek. Meeting room configuration and multiple camera calibration in meeting analysis. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 37–44. ACM, 2005.

[38] Z. Zeng, M. Pantic, G. Roisman, T. S. Huang, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[39] Y. Zhang, L. Zhang, and M. A. Hossain. Adaptive 3d facial action intensity estimation and emotion recognition. *Expert Systems with Applications*, 42(3):1446–1464, 2015.