PLM-IPE: A Pixel-Landmark Mutual Enhanced Framework for Implicit Preference Estimation

Federico Becattini
University of Florence
Italy
federico.becattini@unifi.it

Xuemeng Song
Shandong University
China
sxmustc@gmail.com

Claudio Baecchi
University of Florence
Italy
claudio.baecchi@unifi.it

Shi-Ting Fang
Shandong University
China
fangshiting@mail.sdu.edu.cn

Claudio Ferrari
University of Florence
Italy
claudio.ferrari@unifi.it

Liqiang Nie
Shandong University
China
nieliqiang@gmail.com

Alberto Del Bimbo
University of Florence
Italy
alberto.delbimbo@unifi.it

ABSTRACT

In this paper, we are interested in understanding how customers perceive fashion recommendations, in particular when observing a proposed combination of garments to compose an outfit. Automatically understanding how a suggested item is perceived, without any kind of active engagement, is in fact an essential block to achieve interactive applications. We propose a pixel-landmark mutual enhanced framework for implicit preference estimation, named PLM-IPE, which is capable of inferring the user's implicit preferences exploiting visual cues, without any active or conscious engagement. PLM-IPE consists of three key modules: pixel-based estimator, landmark-based estimator and mutual learning based optimization. The former two modules work on capturing the implicit reaction of the user from the pixel level and landmark level, respectively. The last module serves to transfer knowledge between the two parallel estimators. Towards evaluation, we collected a real-world dataset, named SentiGarment, which contains 3,345 facial reaction videos paired with suggested outfits and human labeled reaction scores. Extensive experiments show the superiority of our model over state-of-the-art approaches.

CCS CONCEPTS

• **Information systems → Sentiment analysis**.

KEYWORDS

Implicit preference estimation, Mutual learning, Facial Reaction Estimation, Sentiment Analysis

ACM Reference Format:
Federico Becattini, Xuemeng Song, Claudio Baecchi, Shi-Ting Fang, Claudio Ferrari, Liqiang Nie, and Alberto Del Bimbo. 2021. PLM-IPE: A Pixel-Landmark Mutual Enhanced Framework for Implicit Preference Estimation. In *ACM Multimedia Asia (MMAsia '21), December 1–3, 2021, Gold Coast, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3469877.3490621

MMAsia '21, December 1–3, 2021, Gold Coast, Australia

ACM ISBN 978-1-4503-8607-4/21/12…$15.00
https://doi.org/10.1145/3469877.3490621

1 INTRODUCTION

Online retailers hinge on the effectiveness of recommendation systems to suggest products to users [4, 9, 12, 28, 30], which not only increases the profit for retailers but also convenience for users. In a sense, such systems become more effective when users create an online profile or when previous purchases and preferences are made available. Even first-time customers can be tracked online via their browsing history as they visit the shop. However, the same cannot be said for physical retailers, where no such data is available even for regular customers. Although some shops have resorted to asking the user to navigate a digital marketplace from inside the shop using a terminal or touchscreen to collect useful cues, they still require certain forms of engagement from the user, who is often not willing to interact or share personal tastes and preferences. This leaves the retailer with the problem of addressing user needs without any prior knowledge or feedback.

In this paper, we address the issue by developing a computer vision-based system capable of inferring user preferences without any active or conscious engagement. We propose a pixel-landmark mutual enhanced framework for implicit preference estimation, named PLM-IPE, which jointly captures the facial reaction at both pixel level and landmark level. Specifically, PLM-IPE consists of three key modules: pixel-based estimator (PBE), landmark-based estimator (LBE), and the mutual learning based optimization. The pixel-based estimator aims to capture the video content at the pixel level, while the landmark-based estimator works on characterizing the user's facial expression more explicitly at the landmark level, where the local movements of his/her facial keypoints are modeled. The mutual learning based optimization module targets at mutually transferring knowledge between the two parallel estimators with the underlying philosophy that there should be some latent consistency between the results of the two estimators. To evaluate our proposed model, we create a dataset, named SentiGarment, by collecting 120 volunteers' facial reactions to the provided fashion outfits. Ultimately, SentiGarment consists of 3,345 samples, each of which comprises a micro-video that contains the volunteer's facial reaction to the given fashion outfit and his/her subjective degree of interest towards the observed outfit.

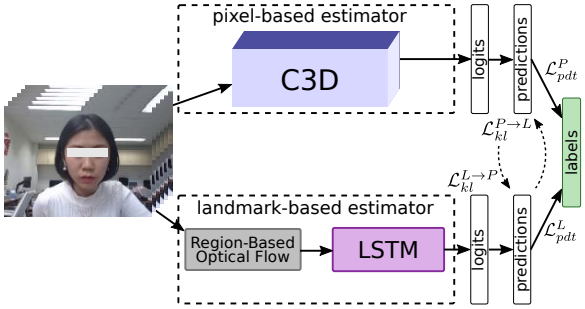

Figure 1: Architecture of our proposed system, comprising three modules: the pixel-based estimator, landmark-based estimator, and the mutual learning based optimization.

The main contributions of this paper are the following:

- We define the task of the visual-based implicit preference estimation, i.e. inferring the degree of interest of a user towards an observed item only through facial expression analysis of the user.
- We propose a pixel-landmark mutual enhanced framework for implicit preference estimation, named PLM-IPE, which comprehensively analyzes the human facial expression from both pixel level and landmark level, and employs mutual learning to boost the model performance.
- We present SentiGarment, a dataset of 3,345 facial-reaction videos with manually annotated user preference labels. Experiments on the dataset prove the effectiveness of our approach.

2 RELATED WORK

The analysis of human emotions based on facial responses is a central computer vision task that has been studied for decades. Both spontaneous and posed facial expressions represent an effective non-verbal means of communication used to convey emotions, and numerous studies [1, 8, 13–15, 18, 19, 24, 29, 35] have proposed different ways to categorize and understand them. For a comprehensive survey, we refer the reader to [17, 23]. Both posed and spontaneous expressions are yet characterized by the fact that they serve to communicate, and so are *voluntary*. A different category of emotional responses is known in literature as micro-expressions [34]. These are subtle and short involuntary facial movements, which are challenging to detect and recognize.

In practice, possible emotional responses are way more complex and can be expressed via a multitude of subtle facial expressions. To represent such a complexity, a different paradigm is being used that models the possible emotional states as continuous values in *valence* and *arousal* space [3]. Valence shows how positive or negative an emotional state is, whilst arousal shows how passive or active it is. Given the applicability of such model in realistic scenarios, several methods are being developed to estimate valence and arousal from images or videos *e.g.* [2, 6, 21, 37]. Simultaneously, researchers are putting a lot of effort in collecting datasets with annotated valence and arousal values to further improve this research direction. One of the largest annotated datasets is the AffWild database [22], which contains a large set of "in the wild" face images with valence and arousal annotations, provided by experts. However, the expressions contained in this kind of datasets are still too strong to resemble an involuntary reaction. Given the slightly perceivable nature of spontaneous reactions, the lack of methods for their estimation is a direct consequence of the absence of proper data.

3 METHOD

3.1 Problem Formulation

As a pioneer study, we treat the implicit preference estimation as a three-class classification. Namely, we only estimate the user's preference as positive, neutral, and negative. Suppose we have a set of N videos $\mathcal{X} = \{x_i\}_{i=1}^N$ that record the users' reactions, where x_i is the i-th video. Each video is associated with its ground truth preference label $\boldsymbol{y}_i \in \mathbb{R}^3$ of the human reaction, which is a 3-D one-hot vector. We aim to develop a vision-based reaction estimator capable of automatically inferring the implicit preference of the user toward an item based on the video showing his/her reaction.

3.2 Pixel-based Estimator

CNN-based approaches play an important role in video analysis [7, 32], since they capture the video content from the detailed pixel level. Due to the remarkable performance of the deep 3-dimensional convolutional networks (3D ConvNets) [31] in this research line, we adopt it to encode the content of the input video x_i and derive its corresponding pixel-based feature $\boldsymbol{f}_i^P$. 3D ConvNets consist of eight convolutional layers and two fully-connected layers, where a respective 3D pooling layer is linked behind the first, second, fourth, sixth and eighth convolutional layers. In particular, following the settings proposed in [31], we split the video x_i into a set of non-overlapped 16-frame clips, and hence re-define the video with the representation of $\mathcal{X}_i \in \mathbb{R}^{l\times w\times h\times c}$, where $l = 16$ is the number of frames, c is the number of channels, h and w are the height and width of the frame, respectively. We then feed the video clips into a 3D ConvNet and use its output, i.e., a 4096-dimensional vector, as the pixel-based feature $\boldsymbol{f}_i^P \in \mathbb{R}^{4096}$ of the input video x_i as follows,

$$\boldsymbol{f}_i^P = \text{3DConvNet}(\mathcal{X}_i). \tag{1}$$

Thereafter, we adopt a fully-connected layer to derive the preference distribution over three categories, i.e., positive, neutral, and negative, of the user in each video as follows,

$$\hat{\boldsymbol{y}}_i^P = softmax(\boldsymbol{W}_p \boldsymbol{f}_i^P + \boldsymbol{b}_p), \tag{2}$$

where $\boldsymbol{W}_p \in \mathbb{R}^{4096\times 3}$ and $\boldsymbol{b}_p \in \mathbb{R}^3$ are the parameters of the fully-connected layer. $softmax(\cdot)$ is the softmax active function.

3.3 Landmark-based Estimator

Along with the pixel-based approach, we rely on a landmark-based approach which is more explicitly related to facial expressions and their underlying semantics. Facial landmarks are a set of keypoints that localize salient regions in a face such as mouth, nose and eyebrows. Such keypoints are capable of describing local deformations of a face simply through their relative positions. Inspired by [25], that exploits facial landmarks to identify Action Units, we build a facial descriptor characterizing local movements of regions connecting the landmarks.

First, for each video x_i we localize the face in each frame j and crop it to remove background noise. We then estimate facial landmarks using [5], which yields a set of 68 2-dimensional keypoints on the image. In order to achieve invariance to scale, translation and rotation, we align each detected face in the video stream using the identified landmarks. Following [25], we then divide the face into 36 regions by connecting adjacent keypoints. Local movements of these regions have been shown to correlate with facial Action Units and can therefore be associated to the emotions of the subject. The rationale of this idea is to obtain a descriptor that carries information about such emotions in order to predict the observed reaction of the user. To obtain the final frame-wise descriptor we compute dense optical flow using [16] and we bin each motion vector into its corresponding facial region, thus generating a motion histogram distributed over the face. We split orientation and magnitude for each vector, obtaining two 36-dimensional descriptors $\boldsymbol{h}_{ij}^O$ and $\boldsymbol{h}_{ij}^M$ for frame j of video x_i. The histograms are then normalized and concatenated into a final 72-dimensional landmark-based descriptor $\boldsymbol{h}_{ij}^L = \boldsymbol{h}_{ij}^O \oplus \boldsymbol{h}_{ij}^M$.

Since the goal is to estimate the implicit preference for the user in each video, we feed the sequence of descriptors for each video into an LSTM-based prediction model. Consequently, we compress the whole sequence of descriptors into a latent representation as,

$$\boldsymbol{f}_i^L = \text{LSTM}(\boldsymbol{h}_{ij}^L). \tag{3}$$

where $\boldsymbol{f}_i^L \in \mathbb{R}^H$ denotes the latent representation of the video x_i, and H is the dimension of the representation.

Similar to the pixel-based estimator, we exploit a linear layer followed by a softmax activation to map the video descriptor $\boldsymbol{f}_i^L$ into a probability distribution $\hat{\boldsymbol{y}}_i^L$ over the preference labels,

$$\hat{\boldsymbol{y}}_i^L = softmax(\boldsymbol{W}_L \boldsymbol{f}_i^L + \boldsymbol{b}_L) \tag{4}$$

with $\boldsymbol{W}_L \in \mathbb{R}^{H\times 3}$ and $\boldsymbol{b}_L \in \mathbb{R}^3$ the learnable parameters of the layer.

3.4 Mutual Learning based Optimization

In this work, we cast the implicit preference estimation as a three-class classification. Accordingly, we have the following objective functions for the pixel-based and landmark-based estimators:

$$\begin{cases} \mathcal{L}_{pdt}^P = \sum_{i=1}^N -\boldsymbol{y}_i \log(\hat{\boldsymbol{y}}_i^P), \\ \mathcal{L}_{pdt}^L = \sum_{i=1}^N -\boldsymbol{y}_i \log(\hat{\boldsymbol{y}}_i^L), \end{cases} \tag{5}$$

where $\boldsymbol{y}_i \in \mathbb{R}^3$ is the ground truth preference of the user in the video x_i. $\hat{\boldsymbol{y}}_i^P \in \mathbb{R}^3$ and $\hat{\boldsymbol{y}}_i^L \in \mathbb{R}^3$ are the predicted preference distribution vectors of the human reaction in video x_i from the pixel-based and landmark-based estimators, respectively.

Moreover, although the two estimators model the human reactions from different perspectives, for the same video, their evaluations should still be somehow consistent. Therefore, we incorporate the mutual learning strategy [33, 36], which has shown remarkable performance in knowledge distillation between two learners, to encourage information sharing with each other. In particular, we adopt the most popular Kullback-Leibler divergence between $\hat{\boldsymbol{y}}_i^L$ and $\hat{\boldsymbol{y}}_i^P$ to encourage the consistency between the two learners. Specifically, we define the objective function as follows,

$$\begin{cases} \mathcal{L}_{kl}^{L\to P} = \sum_{i=1}^N KL(\hat{\boldsymbol{y}}_i^L || \hat{\boldsymbol{y}}_i^P) = \sum_{i=1}^N \hat{\boldsymbol{y}}_i^L \log \frac{\hat{\boldsymbol{y}}_i^L}{\hat{\boldsymbol{y}}_i^P}, \\ \mathcal{L}_{kl}^{P\to L} = \sum_{i=1}^N KL(\hat{\boldsymbol{y}}_i^P || \hat{\boldsymbol{y}}_i^L) = \sum_{i=1}^N \hat{\boldsymbol{y}}_i^P \log \frac{\hat{\boldsymbol{y}}_i^P}{\hat{\boldsymbol{y}}_i^L}, \end{cases} \tag{6}$$

where $L \to P$ and $P \to L$ denote the knowledge transferring from landmark-based estimator to pixel-based estimator and its opposite, respectively. $\mathcal{L}_{kl}^{L\to P}$ and $\mathcal{L}_{kl}^{P\to L}$ refer to the regularization for the landmark-based and pixel-based methods, respectively.

Ultimately, our final objective function can be formulated as,

$$\begin{cases} \mathcal{L}^L = \mathcal{L}_{pdt}^L + \mathcal{L}_{kl}^{L\to P}, \\ \mathcal{L}^P = \mathcal{L}_{pdt}^P + \mathcal{L}_{kl}^{P\to L}. \end{cases} \tag{7}$$

Overall, we alternatively optimize the landmark-based and pixel-based estimators with the above losses. Notably, for each estimator optimization, only the corresponding parameters need to be optimized. Once the whole network gets well-optimized, we estimate the overall score of human reactions for a given video as follows,

$$\hat{\boldsymbol{y}}_i = \lambda \hat{\boldsymbol{y}}_i^L + (1-\lambda)\hat{\boldsymbol{y}}_i^P, \tag{8}$$

where λ is a trade-off parameter to balance the two parts.

4 EXPERIMENTS

4.1 Dataset

Although some public datasets such as CK+ [26] and AFEW 8.0 [11] are available to estimate human emotions, they fail to capture human reactions toward the items. Therefore, we constructed our own dataset by inviting 120 volunteers. In particular, we created a specific web application that is able to both collect the volunteer's demographic attributes and allow the volunteer to give his/her personal rate (ranges from 0 to 100) on each observed outfit. Moreover, while observing the outfits, the volunteer's reaction is captured using the volunteer's device camera. While being recorded the volunteer is free to observe the outfit for a maximum time of 10 seconds. In each annotation session, 15 outfits are proposed. Ultimately, we obtained the final dataset, named SentiGarment, comprising 3,345 pieces of data of human facial reactions towards fashion outfits. Each piece of data consists of a top and a bottom image, a video with the volunteer's reaction to the top-bottom pair, his/her preference score and personal data.

We split the dataset into train and test set following the ratio of 8 : 2 based on the number of volunteers. We first separate the testing set selecting 27 volunteers so that: *i)* there is no overlap of these volunteers with the training set, *ii)* each volunteer has participated in the annotation process just once (i.e. all volunteers in the testing set have the same number of videos), and *iii)* we try to balance the male/female ratio as best as possible. With these constraints, the final testing set consists of 180 and 225 videos from 12 females and 15 males, respectively. Notably, since the same user can participate more than 1 session, there is a sensible redundancy in the training set. Based on the reaction scores, we assign each video in the dataset to one of three classes: *positive*, *neutral*, and *negative*. Finally, we obtained 1,334 positive videos, 842 neutral videos, and 1,169 negative videos.

Table 1: Performance comparison among different methods.

Method	Accuracy(%)
PBE	49.14
LBE	40.39
EarlyFusion	49.14
LateFusion	50.37
AffWildNet	41.73
MIMAMO-Net	45.43
Ours	**52.84**

4.2 Implementation Details

For the pixel-based estimator, we follow the experiment setting of [31]. The input dimension is set as $16 \times 112 \times 112 \times 3$, i.e., $l = 16$, $w = 112$, $h = 112$, and $c = 3$. The 3D ConvNet has 8 convolution layers, 5 max pooling layers and 2 fully connected layers. All 3D convolution kernels are $3 \times 3 \times 3$ with stride $1 \times 1 \times 1$. All pooling kernels are $2 \times 2 \times 2$ with stride $2 \times 2 \times 2$, except for the first pooling layer which has kernel size of $1 \times 2 \times 2$ and stride $1 \times 2 \times 2$. All fully-connected layers have 4096 output units. Specifically, we adopted the adaptive moment estimation method (Adam) [20], and set the initial learning rate to $1e^{-4}$. For the landmark-based estimator, we set the dimension of the latent representation yielded by the landmark-based estimator, i.e., the number of hidden states in LSTM, $H = 256$. In addition, we utilized the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a fixed learning rate of $5e^{-3}$. The batch size is set to 32 and the model is fine-tuned for 500 epochs. We empirically set the trade-off and temperature parameters as 0.3 and 10, respectively. All experiments are implemented by PyTorch.

4.3 On Model Comparison

To justify our proposed method, we adopted the following methods.

- **PBE**. This baseline estimates the user's preference only relying on the pixel-based estimator, which can be easily derived from our proposed method by only optimizing the $\mathcal{L}^P_{pdt}$ in Eqn.(5).
- **LBE**. Similarly, LBE only uses the landmark-based estimator, and can be derived by only optimizing the $\mathcal{L}^L_{pdt}$ in Eqn.(5).
- **EarlyFusion.** We jointly trained the pixel-based and landmark-based estimators by feeding the weighted sum of their predicted preference distributions, i.e., $\hat{\mathbf{y}} = \beta_1 \hat{\mathbf{y}}^L_t + \beta_2 \hat{\mathbf{y}}^L_t$, into the cross-entropy loss. For testing, we used $\hat{\mathbf{y}}$ as the result.
- **LateFusion.** We first trained the pixel-based and landmark-based estimators separately. Once each estimator is well-trained, we used the weighted sum of their respective predicted results as the final estimated preference distribution of the given video.
- **AffWildNet**[22]. This baseline is composed of a backbone CNN (either ResNet-50 or VggFace [27]), upon which two Gated Recurrent Units (GRU) are stacked to regress valence and arousal. We fine-tuned the pre-trained AffWildNet on our dataset by substituting the regression layer with a classification one.
- **MIMAMO-Net**[10]. This method uses a two-stream recurrent network to combine the micro- and macro-motion features to improve video emotion recognition. Following the original work, we aligned and extracted faces from our videos using the OpenFace toolkit, and then we extracted the respective features from the *pool5_7x7_s1* layer. Finally, we replaced the regressor with a classifier to accommodate our task.

Table 2: The ablation study of our proposed method.

Method	Accuracy(%)
PBE	49.14
LBE	40.39
PBE-w/-Mut	49.14
LBE-w/-Mut	41.48

Table 1 shows the comparison between different approaches on the SentiGarment dataset. We can derive the following observations: 1) our method surpasses all baselines, which shows the superiority of the framework over existing methods. 2) PBE shows superiority over LBE, which implies that employing a 3D ConvNet to model the video content is more powerful than only utilizing the landmarks and optical flow. One possible reason is that only relying on the landmarks may lead to the loss of some useful cues contained in the video. And 3) our method outperforms both EarlyFusion and LateFusion, reflecting the superiority of utilizing the mutual learning strategy to seamlessly combine the pixel-based classifier and the landmark-based classifier.

4.4 On Mutual Learning

To get a thorough understanding of the mutual learning based optimization in our model, we conducted experiments to learn the effect of mutual learning for our PBE and LBE. In particular, we introduced two derivatives of our model: **PBE-w/-Mut** and **LBE-w/-Mut**, where we train our proposed framework, but only used the $\hat{\boldsymbol{y}}^P_i$ and $\hat{\boldsymbol{y}}^L_i$ as the predicted results of PBE-w/-Mut and LBE-w/-Mut, respectively. Table 2 shows the ablation study results. As can be seen, PBE-w/-Mut outperforms PBE, which suggests that the mutual learning strategy is capable of transferring useful knowledge from LBE to PBE. However, we noticed that the result of PBE-w/-Mut is the same as that of PBE. The reason may be that PBE is much powerful than LBE, and could not distill additional useful knowledge from LBE during the mutual learning process.

5 CONCLUSION

We presented a novel computer-vision based implicit preference estimation pipeline, based on three modules: pixel-based estimator, landmark-based estimator and mutual learning based optimization. The approach is a first step towards building an interactive recommendation system. By relying only on computer vision, the method is non-intrusive and does not require active engagement of the customer. To evaluate such method, we collected a dataset of facial reactions paired with preference labels and trained different models. We showed the benefits of exploiting the mutual learning based optimization to combine the pixel-based estimator and the landmark-based estimator. In the future, we plan to investigate the interactive recommendation systems based on our current work.

ACKNOWLEDGMENTS

This work was partially supported by the Italian MIUR within PRIN 2017, Project Grant 20172BH297: I-MALL - improving the customer experience in stores by intelligent computer vision; the Shandong Provincial Natural Science Foundation, No.:ZR2019JQ23; the Key R&D Program of Shandong (Major scientific and technological innovation projects), No.:2020CXGC010111.

REFERENCES

[1] Luigi Ariano, Claudio Ferrari, Stefano Berretti, and Alberto Del Bimbo. 2021. Action Unit Detection by Learning the Deformation Coefficients of a 3D Morphable Model. *Sensors* 21, 2 (2021), 589.

[2] Claudio Baecchi, Tiberio Uricchio, Marco Bertini, and Alberto Del Bimbo. 2017. Deep sentiment features of context and faces for affective video analysis. In *Proceedings of the ACM International Conference on Multimedia Retrieval*. ACM, 72–77.

[3] Lisa Feldman Barrett. 1998. Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition & Emotion* 12, 4 (1998), 579–599.

[4] Wolmer Bigi, Claudio Baecchi, and Alberto Del Bimbo. 2020. Automatic Interest Recognition from Posture and Behaviour. In *Proceedings of the ACM International Conference on Multimedia*. Association for Computing Machinery, 2472–2480.

[5] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1021–1030.

[6] Wei-Yi Chang, Shih-Huan Hsu, and Jen-Hsien Chien. 2017. FATAUVA-Net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 17–25.

[7] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. Micro Tells Macro: Predicting the Popularity of Micro-Videos via a Transductive Model. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 898–907.

[8] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. 2017. Learning spatial and temporal cues for multi-label facial action unit detection. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 25–32.

[9] Lavinia De Divitiis, Federico Becattini, Claudio Baecchi, and Alberto Del Bimbo. 2021. Style-Based Outfit Recommendation. In *Proceedings of the International Conference on Content-Based Multimedia Indexing*. IEEE, 1–4.

[10] Didan Deng, Zhaokang Chen, Yuqian Zhou, and B. Shi. 2020. MIMAMO Net: Integrating Micro- and Macro-motion for Video Emotion Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 2621–2628.

[11] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. 2018. EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction. In *Proceedings of the ACM International Conference on Multimodal Interaction*. ACM, 653–656.

[12] Lavinia De Divitiis, Federico Becattini, Claudio Baecchi, and Alberto Del Bimbo. 2020. Garment Recommendation with Memory Augmented Neural Networks. In *Proceedings of the International Conference on Pattern Recognition Worlshops*. Springer, 282–295.

[13] Shichuan Du, Yong Tao, and Aleix M Martinez. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences* 111, 15 (2014), E1454–E1462.

[14] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17, 2 (1971), 124–129.

[15] Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[16] Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the Scandinavian Conference on Image Analysis*. Springer, 363–370.

[17] Beat Fasel and Juergen Luettin. 2003. Automatic facial expression analysis: a survey. *Pattern recognition* 36, 1 (2003), 259–275.

[18] Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, and Alberto Del Bimbo. 2017. A dictionary learning-based 3D morphable shape model. *IEEE Transactions on Multimedia* 19, 12 (2017), 2666–2679.

[19] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. 2012. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* 109, 19 (2012), 7241–7244.

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net.

[21] Dimitrios Kollias, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. 2017. Recognition of affect in the wild using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 26–33.

[22] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. 2019. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision* 127, 6 (2019), 907–929.

[23] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE Transactions on Afective Cmputing* (2020), 1–20.

[24] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2019. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 10924–10933.

[25] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, Guoying Zhao, and Xiaolan Fu. 2015. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transactions on Affective Computing* 7, 4 (2015), 299–310.

[26] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 94–101.

[27] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. British Machine Vision Association, 1–12.

[28] Dikshant Sagar, Jatin Garg, Prarthana Kansal, Sejal Bhalla, Rajiv Ratn Shah, and Yi Yu. 2020. PAI-BPR: Personalized Outfit Recommendation Scheme with Attribute-wise Interpretability. In *Proceedings of the IEEE International Conference on Multimedia Big Data*. IEEE, 221–230.

[29] Georgia Sandbach, Stefanos Zafeiriou, and Maja Pantic. 2012. Local normal binary patterns for 3D facial action unit detection. In *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 1813–1816.

[30] Xuemeng Song, Xianjing Han, Yunkai Li, Jingyuan Chen, Xin-Shun Xu, and Liqiang Nie. 2019. GP-BPR: Personalized Compatibility Modeling for Clothing Matching. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 320–328.

[31] Du Tran, Lubomir D. 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*. IEEE.

[32] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoquan Chen. 2019. Neural multimodal cooperative learning toward micro-video understanding. *IEEE Transactions on Image Processing* 29 (2019), 1–14.

[33] Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. 2021. Comprehensive Linguistic-Visual Composition Network for Image Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1369–1378.

[34] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. 2013. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior* 37, 4 (2013), 217–230.

[35] Huiyuan Yang, Lijun Yin, Yi Zhou, and Jiuxiang Gu. 2021. Exploiting Semantic Embedding and Visual Feature for Facial Action Unit Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 10482–10491.

[36] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2018. Deep Mutual Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4320–4328.

[37] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, and Shiguang Shan. 2020. M3F: Multi-Modal Continuous Valence-Arousal Estimation in the Wild. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 632–636.