A Psychologically-Inspired Match-Score Fusion Model for Video-Based Facial Expression Recognition

Albert Cruz, Bir Bhanu, Songfan Yang,

VISLab, EBUII-216, University of California Riverside,
Riverside, California, USA, 92521-0425
{acruz, bhanu, syang}@ee.ucr.edu

Abstract. Communication between humans is rich in complexity and is not limited to verbal signals; emotions are conveyed with gesture, pose and facial expression. Facial Emotion Recognition and Analysis (FERA), the set of techniques by which non-verbal communication is quantified, is an exemplar case where humans consistently outperform computer methods. While the field of FERA has seen many advances, no system has been proposed which scales well to very large data sets. The challenge for computer vision is how to automatically and non-heuristically downsample the data while maintaining a minimum representational power that does not sacrifice accuracy. In this paper, we propose a method inspired by human vision and attention theory [2]. Video is segmented into temporal partitions with a dynamic sampling rate based on the frequency of visual information. Regions are homogenized by an experimentally selected match-score fusion technique. The approach is shown to increase classification rates by over baseline with the AVEC 2011 video-subchallenge [15].

Keywords: vision and attention theory; avatar image registration; local phase quantization

1 Introduction

The field of video-based emotion recognition has been an active area of work [13, 14, 23] and has progressed with the help of standardized data sets such as JAFFE [10], among the earliest, and state-of-the-art data sets such as Cohn-Kanade+ [9] and the MMI Database [19]. However, despite the existence of these standards, many papers select subsets of data or do not detail how training and testing sets are generated [18]. Challenge data sets such as FERA Challenge 2011 [18] and The Audio/Visual Emotion Challenge and Workshop (AVEC 2011) [15] provide a benchmark to compare different emotion recognition systems on common ground.

Facial Expression Recognition and Analysis (FERA) typically extract frontal face images and compute either appearance features (such as Gabor wavelets [22] and the family of LBP type features [7]) or geometric relationship features (such as AAM-

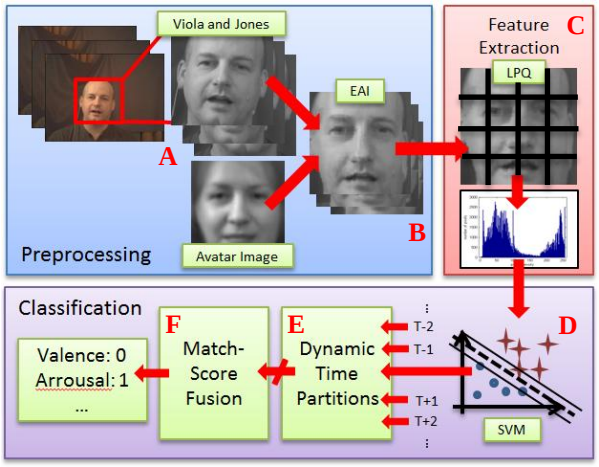

Fig. 1. System overview.

based features [9]). Facial expression is commonly detected in terms of Facial Action Units, a minimal set of facial muscle actions, or the "big six" emotion labels. Fontaine *et al.* [3] assert that basic emotions can be described along the four dimensions: activation, expectancy, power and valence. AVEC 2011 uniquely requires detection of emotion along these four axes.

Approaches commonly assume that video sequences are pre-cut such that a subject expresses a single emotion over the sequence, or take advantage of labeling the time point where emotion is most intense, known as the apex. While knowledge of apex location positively affects performance, labeling requires an expert. In uncut interview footage such as in Solid-SAL [11]—the data used in AVEC 2011—subjects express multiple apex maxima in a single video. An automatic approach for uncut and unlabeled video that does not require apex labeling is proposed.

1.1 Vision and Attention Theory Inspiration

Humans can suddenly change emotion state, meriting high frame-rates for precise detection of changes. However, a constantly high frame-rate may not be necessary for the whole video. A subject is not likely to change emotion state while idle. High frame rates unnecessarily increase sample size. Intuitively, a recognition system should devote fewer resources when the subject in video is idle and more resources when there is action in the video, e.g., a high frame rate is required to properly describe an animated or speaking subject to represent changes in visual information whereas only a few frames are needed to describe a subject when idle with little change in visual information. In the human visual system (HVS), steady state visual information is processed at a rate of <1Hz [2]. This rate increases proportionally with the frequency of visual stimulus [5]. In this paper, we use this concept as inspiration for partitioning videos into temporal segments of varying sampling rates which increase proportionally with the frequency of visual information. Visual stimulus is quantified using a Discrete-Time Fourier-Transform (DTFT) of motion features and the dominant frequency controls the sampling rate locally for each time-segment.

Contribution. A perceptual psychologically-inspired model for segmenting video into time partitions with a dynamic, minimal frame rate needed to meaningfully represent each local volume is proposed. We propose homogenizing dynamic time partitions with a combination-based match-score fusion technique which is experimentally selected from performance on the development set. This approach is robust in that it does not require apex labeling and does not make assumptions that would not scale to real world data. This approach directly addresses the sample dimensionality problem in AVEC 2011 while reducing loss in precision of emotion state changes.

2 Technical Approach

We propose the following recognition pipeline given in Fig. 1: (A) Frontal face region of interest are detected with a Viola-Jones framework [20]. (B) Faces are registered using Avatar Image Registration [21] (C) Local Phase Quantization features (LPQ) [12] features are extracted and (D) each frame is classified using a linear Support Vector Machine (SVM). (E) Video data is segmented into time partitions of varying sampling rate from 1Hz to the maximum video frame rate, controlled by the dominant frequency of the DTFT of motion features. (F) Test results are fused at the combination match-score level [6] using multiple matchers locally for each partition.

2.1 Face Detection and Registration

Viola and Jones Face Detection. A Viola-Jones detector [20] is trained to detect frontal faces in order to extract regions of interest. In the Solid-SAL data set**,** pose may be extreme such that a frontal face is obscured, or a face not totally in frame. For training, frames where the detector does not detect a frontal face are treated as bad samples and are withheld from model training. For testing, when the detector fails, it is not possible to properly classify the result using a model trained on frontal face features, as the image is not a frontal face and the result would be incorrect. In testing cases where features are not available, we assign the label with the highest *a priori* probability from training, for that emotion.

Avatar Image Registration. Images are aligned with Avatar Image Registration [21] which aligns face images at the scene-level with SIFT Flow [8]. Each frame is spatially warped to a single reference frame, titled the Avatar Reference; enforcing global alignment. This powerful but simple approach will hallucinate a frontal, non-posed face image by aligning facial structures across individuals and emotions while maintaining internal-structure information.

SIFT Flow. SIFT Flow addresses the problem of image registration by aligning a query image to a target image at the scene level by spatially warping the query image to match the target image. The alignment task is formulated as an optical flow problem using dense SIFT descriptors. The objective function for SIFT Flow is

formulated as:

$$E(\boldsymbol{w}) = \sum_{\boldsymbol{p}} \min(\|s_A(\boldsymbol{p}) - s_i(\boldsymbol{p}+\boldsymbol{w})\|_1) + \sum_{\boldsymbol{p}} u^2(\boldsymbol{p}) + v^2(\boldsymbol{p})/\sigma^2 + \sum_{(\boldsymbol{p},\boldsymbol{q})\epsilon\, N_{\boldsymbol{q}}} (\min(\alpha|u(\boldsymbol{p}) - u(\boldsymbol{q})|) + \min(\alpha|v(\boldsymbol{p}) - v(\boldsymbol{q})|)) \quad (1)$$

where $\boldsymbol{p} = (x, y)$ is a pixel in the image, $\boldsymbol{w}(\boldsymbol{p})$ is the motion vector at pixel $\boldsymbol{p}$ between the query and target images where $\boldsymbol{w}(\boldsymbol{p}) = (u(\boldsymbol{p}), v(\boldsymbol{p}))$, s_A and s_i are the dense SIFT descriptors of the target image and the query image i respectively and ε is the 4-member neighborhood about $\boldsymbol{p}$. One of the expressed purposes of this algorithm is to counteract object motion between two scenes. This makes SIFT Flow well suited to correcting facial plane motion.

Avatar Reference. While the SIFT Flow algorithm is powerful enough to align two images, a target image must be automatically generated for each frame to be registered to. There exists some veridical frontal face image $\hat{A}$, the Avatar Reference, that consists of deterministic values. Observed pixels in the data set are samples from the Avatar Image that are corrupted by a zero-mean noise as a result of extraction errors, facial motion, skin tone, etc. Let f_i be an image. A pixel is observed as:

$$f_i(\boldsymbol{p}) = \hat{A}(\boldsymbol{p}) + \eta \quad (2)$$

where η is zero-mean noise. The minimum variance unbiased solution to Eq. (2) is:

$$\tilde{A}(\boldsymbol{p}) = \frac{1}{N}\sum_i f_i(\boldsymbol{p}) \quad (3)$$

where f_i is the ith image in data and N is the total number of frames in the data.

2.2 Feature Extraction

LPQ was originally introduced by Ojansivu and Heikkila [12] as a blur-invariant texture descriptor. In this paper, LPQ is preferred over LBP because the averaging from Eq. (3) causes some loss of high-frequency information. Images are reduced to 8×8 regions and LPQ features are generated for each region.

Local Phase Quantization. In step one of LPQ, the 2-D DFT is computed locally for four coefficients. Let f be a sample image, and u be the image frequency coefficient. LPQ computes the local Fourier transform as:

$$F(\boldsymbol{u}, \boldsymbol{x}) = \sum_{\boldsymbol{y}\epsilon N_x} f\,(\boldsymbol{x} - \boldsymbol{y}) e^{-j2\pi \boldsymbol{u}^t \boldsymbol{y}} \quad (4)$$

where N_x is the neighborhood about $\boldsymbol{x}$. F can be rewritten in terms of matrixes: $\boldsymbol{F} = \boldsymbol{w}_{\boldsymbol{u}}^{t}\boldsymbol{f}(\boldsymbol{x})$, where $\boldsymbol{w}_{\boldsymbol{u}}$ is the basis vector for a given frequency $\boldsymbol{u}$. LPQ computes the four complex frequencies: $\boldsymbol{u} = \{[\alpha, 0]^t, [0, \alpha]^t, [\alpha, \alpha]^t, [\alpha, -\alpha]^t\}$. In step three, $\boldsymbol{F}$ is decorrelated with a whitening transform and the values are quantized. Let $\boldsymbol{G}$ be the SVD of $\boldsymbol{F}$, and $\langle \boldsymbol{G} \rangle_x$ be the x-th component of $\boldsymbol{G}$ corresponding to the decorrelated 2-D DFT of neighborhood N_x from Eq. (4). $\langle \boldsymbol{G} \rangle_x$ is quantized with:

$$\langle \boldsymbol{G} \rangle_x = \begin{cases} 1 & \text{if } \langle \boldsymbol{G} \rangle_x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$\langle \boldsymbol{G} \rangle_x$ is encoded to an 8-bit scalar—each bit corresponding to one of the neighbors of x— with the following equation:

$$b(x) = \sum_{i=1}^{8} \langle \boldsymbol{G} \rangle_x 2^{i-1} \tag{6}$$

Finally, a 256-bin histogram of $b(x)$ is generated and this is taken to be the feature vector for the image region.

2.3 Psychologically-Inspired Dynamic Temporal Volumes

Three conclusions can be drawn from inspection of the data sets: (A) The frame rate is too high to use all frames for training and testing. Video data in SolidSAL has a frame rate of ~50 FPS resulting in roughly 500,000 frames for the development set. Loading a feature vector of double precision numbers with that many samples and the feature cardinality from Sec. 2.2 requires more than 65GB of memory. (B) SolidSAL videos are interview footage, and there are times where the subject in frame is idle. It is not necessary to have such a high frame rate for these times. (C) States of emotion form homogenous regions. While a subject may change emotion suddenly, changes of emotion state are consistently low frequency, not erratic. From these observations, we propose using vision and attention theory to generate partitions with sampling rate defined dynamically per volume, and fusing the information in each partition with match-score fusion to resist high frequency changes of emotion state.

Intelligent Sampling with Vision and Attention Theory. An immediate solution to the first problem is to sample the data at a constant rate. However, sampling the data too sparsely results in a weak model; too densely, the model is too expensive. It may not be necessary to sample the data at such a high frame rate during the time points when a subject is not active. However, subjects can suddenly change their emotion state. This merits a minimum sampling rate in order to precisely detect change in emotion state. A method is needed for automatically partitioning the data into meaningful segments that capture sudden changes, while weighing idle sequences less. We propose an approach which is perceptual psychology-inspired, where the local sampling rate increases proportionally with the frequency of visual information, see Sec. 1.1.

In this paper, we segment the data into temporal regions of 1Hz. This is inspired by the minimum bound of the HVS [2]. Let τ define the domain of data with which classification is performed for a given temporal volume:

$$\tau = \{t : (t - \beta) < t < (t + \beta)\} \tag{7}$$

where β is the range with of data about t, controlling the sampling rate. When a person is active, the windows size β should be small; a person is idle, β should be large. A problem is posed where β must be selected dynamically. This must be done automatically per frame, without any knowledge of emotion labels. Visual information must be quantified in a way where frequency increases proportionally with activity in the image. SIFT Flow computes motion spatially, not temporally, and is unsuitable for this task. In this paper, we compute motion features as the magnitude of phase based optical flow [4]. Let $v(t)$ be the visual feature content signal computed as:

$$v(t) = \sum_{\boldsymbol{p}} \|g(f_t(\boldsymbol{p}), f_{t-1}(\boldsymbol{p}))\| \tag{8}$$

where $\|g(f_1, f_2)\|$ is the magnitude of optical flow between frames f_1 and f_2. An assumption is made that the frequency of $v(t)$ increases when the subject is active. For these sequences, $v(t)$ is a signal of fluctuating frequency, see Fig. 2. However, when the subject is idle, the signal is a constant level. The dominant frequency of V defines the sampling rate β with the following equation:

$$\beta = \left\lceil \frac{1}{2} \text{argmax}_{\boldsymbol{\omega}}(V(\boldsymbol{\omega})) \right\rceil \tag{9}$$

where $V = \mathcal{F}\{v\}$ computed with the FFT. Psuedo-code for computing $\boldsymbol{\beta}$ is provided in Algorithm 1. $\boldsymbol{\beta}$ is computed once for each temporal partition. For testing, frames which are not sampled are assigned the label given to the partition from the sampled frames.

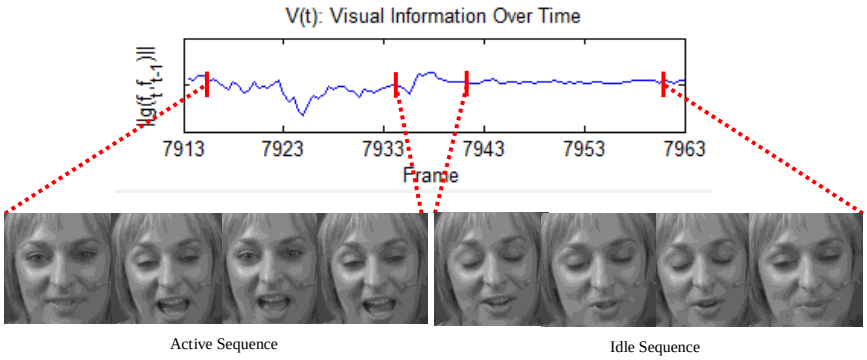

Fig. 2. $\|\boldsymbol{g}(\boldsymbol{f_t}(\boldsymbol{p}), \boldsymbol{f_{t-1}}(\boldsymbol{p}))\|$ for development video 26 about frame 7938.

Algorithm 1. Dynamic assignment of $\boldsymbol{\beta}$ using dominant frequency of $\boldsymbol{V}(\boldsymbol{\omega})$.

procedure computebeta $(\{\boldsymbol{F}_{\boldsymbol{p}}^{1}, \boldsymbol{F}_{\boldsymbol{p}}^{2}, \ldots, \boldsymbol{F}_{\boldsymbol{p}}^{n}\}, \ell)$
Input: Sequence $\{\boldsymbol{F}_{\boldsymbol{p}}^{1}, \boldsymbol{F}_{\boldsymbol{p}}^{2}, \ldots, \boldsymbol{F}_{\boldsymbol{p}}^{n}\}$ where $\boldsymbol{p}$ is a pixel and n is the number of frames; spacing parameter ℓ selected s.t. i iterates 1/s over video sequence.
Output: Sampling rate vector $\boldsymbol{\beta}$ of length n/ℓ.

$\boldsymbol{G}_{\boldsymbol{p}}^{1} = 0$
for $i = 2 \ldots n$:
  $\boldsymbol{G}_{\boldsymbol{p}i} = \text{optical\_flow}(\boldsymbol{F}_{\boldsymbol{p}}^{i-1}, \boldsymbol{F}_{\boldsymbol{p}}^{i})$
$c = 1$
for $j = 1{:}\ell{:}\, n$:
  for $k = \max(\{1, (j-\ell/2)\}) \ldots \min(\{(j+\ell/2), n\})$:
    $v(k) = L^{1}\{\|\boldsymbol{G}_{\boldsymbol{p}k}\|\}$
  $V = \mathcal{F}\{v - \mu_{v}\}$
  $\boldsymbol{\beta}(c) = \lceil \text{argmax}(V(\boldsymbol{\omega}))/2 \rceil$
  $c = c + 1$
return $\boldsymbol{\beta}$

Local Partition Match-Score Fusion. Observation (3) motivates combining match scores locally in each partition to enforce homologous label assignment. In this paper, we propose multiple snapshots, taken from the sampled frames of a local partition, to provide a more robust classification scheme. Fusion is performed at the combination based match-score level [16]. In combination based match-score fusion, the scores, or posterior probabilities from different match-scores are weighted and combined to give a final, scalar match-score. Let $\boldsymbol{X}_{\tau}$ be the sample feature vector of a local time partition τ; let $\widetilde{w}$ be the assigned label from one of the classes $\{w_1, \ldots, w_m\}$; let $p_{\tau}(w_j|\langle \boldsymbol{X}_{\tau} \rangle_i)$ be the output of matcher given data at time i. The classification rule in a match-score combination approach is:

$$\widetilde{w} = \text{argmax}_{\text{j}}\, K\left(P_1(w_j|\langle \boldsymbol{X}_{\tau} \rangle_{t-\beta}), \ldots, P_n(w_j|\langle \boldsymbol{X}_{\tau} \rangle_{t+\beta})\right) \tag{10}$$

$K(.)$ is an aggregator that can implemented with the following rules:

Sum and Product Rules. The aggregators for the sum and product rules are as follows:

$$K\left(P_1(w_j|\langle \boldsymbol{X}_{\tau} \rangle_{t-\beta}), \ldots, P_n(w_j|\langle \boldsymbol{X}_{\tau} \rangle_{t+\beta})\right) = \frac{1}{Z}\sum\nolimits_{i} P(w_j|\langle \boldsymbol{X}_{\tau} \rangle_i) \tag{11}$$

$$K\left(P_1(w_j|\langle \boldsymbol{X}_{\tau} \rangle_{t-\beta}), \ldots, P_n(w_j|\langle \boldsymbol{X}_{\tau} \rangle_{t+\beta})\right) = \prod\nolimits_{i} P(w_j|\langle \boldsymbol{X}_{\tau} \rangle_i) \tag{12}$$

where **(11)** and **(12)** are the weighted sum and product rules, respectively. Z is a normalization s.t. Eq. **(11)** ϵ [0,1].

Extrema Rules. In extrema rules, the class label w is assigned using a min or a max operator:

$$K\left(P_1\left(w_j|\langle \boldsymbol{X}_\tau\rangle_{t-\beta}\right),\ldots,P_n\left(w_j|\langle \boldsymbol{X}_\tau\rangle_{t+\beta}\right)\right)=\min_i\left(P_1\left(w_j|\langle \boldsymbol{X}_\tau\rangle_i\right)\right) \quad \textbf{(13)}$$

$$K\left(P_1\left(w_j|\langle \boldsymbol{X}_\tau\rangle_{t-\beta}\right),\ldots,P_n\left(w_j|\langle \boldsymbol{X}_\tau\rangle_{t+\beta}\right)\right)=\max_i\left(P_1\left(w_j|\langle \boldsymbol{X}_\tau\rangle_i\right)\right) \quad \textbf{(14)}$$

3 Results

Avatar Image Registration is iterated for three generations, which has been shown to be a good tradeoff between computational time and accuracy for LPQ features [21]. All matchers are Support Vector Machines with a linear kernel from the LibSVM MATLAB toolbox [1]. Feature is vector min/max normalized to $[-1,1]$. For a detailed explanation of the data, and the development, training and testing sets, please refer to the Schuller *et al.* [15]. In this paper, we consider only the video sub-challenge. Testing labels are assigned with a model only training fold data.

Results on the development set are given in Table 1. Match-score fusion rule used for testing is determined experimentally using 4-fold random cross-validation with a 75/25 split on development data to avoid overfitting. Similarly to baseline, expectancy and power features are more difficult to detect versus activation and valence. The max rule gives a better average versus other rules and is used for final results on the testing fold for the video sub-challenge.

Table 1. Development Classification Rates on the Video-Subchallenge for Different Match-score Rules.

Accuracy (%)	Activation	Expectancy	Power	Valence	Average
Min Rule	64.48±5.4	65.22±7.9	58.55±6.0	64.21±6.7	63.12
Max Rule	69.30±3.0	65.58±5.5	59.87±6.0	67.79±4.9	65.64
Product Rule	64.15±3.5	67.02±4.8	58.27±5.2	64.71±4.8	63.53
Sum Rule	66.10±2.6	63.70±4.7	59.65±5.1	64.38±5.5	63.45

Table 2. Video-Subchallenge Testing Classification Rates

	Activation		Expectancy		Power		Valence	
Accuracy (%)	WA	UA	WA	UA	WA	UA	WA	UA
Testing	**56.51**	56.87	**59.67**	55.11	**48.52**	49.36	**59.24**	56.72

Results on the video-subchallenge are given in Table 2. WA stands for weighted accuracy, and is the classification rate. UA stands for unweighted accuracy, and is the average recall over the two classes. Proposed approach differs greatly from the baseline in three ways. In the event of Viola and Jones failure the method defaults to *a priori* probabilities—some videos in training and testing folds had a majority of

irretrievable faces, making it a more difficult problem than the development set. The sampling rate is assigned dynamically, and reduces the amount of samples when the subject is not active. This allows training of a model on a larger number of samples without entering the domain of being too computationally expensive. The number of frames retained by our approach is given in Table 3. Whereas approaches similar to the baseline approach might treat each frame in a video independently, in our model a video is comprised of dynamically sized partitions where a label consensus is reached with fusion.

Table 3. Frames retained by Vision and Attention Theory

Set	Total Frames(Fr)	Retained Frames(Fr)
Development	449074	27412
Training	501277	30076
Testing	140125	8383

5 Conclusion

The concept of partitioning emotion video into dynamically sampled segments is explored. A model is proposed for controlling sampling rate in each local partition which was inspired by perceptual psychology where a partition was sampled according to the frequency of visual content. Then, posterior probabilities are combined for each local partition using a combination-based match-score fusion technique. Match-score rule was selected from 4-fold cross validation on the Development set. Performance on testing data is improved via intelligently selecting frames without unnecessarily increasing sample size.

Acknowledgments. Support for this work was provided by NSF IGERT: Video Bioinformatics Grant DGE 0903667. The authors would like to thank the organizers of AVEC 2011 for conducting the challenge.

References

1. Chang, C., & Lin, C. LibSVM: A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/˜cjlin/libsvm
2. Findlay, J., & Gilchrist, I.: Active Vision: The Psychology of Looking and Seeing. Oxford University Press, Oxford (2003)
3. Fontaine, J., Scherer, K., B.R., & Ellsworth, P.E.: The World of Emotions Is Not Two-dimensional. Psychological Science. 18(2), 1050—1057 (2007)
4. Gautama, T., & Van Hulle, M: A Phase-Based Approach to the Estimation of the Optical Flow Field Using Spatial Filtering. Neural Nets, IEEE Trans. on. 13(5), 1127—1136 (2002)
5. Haber, R., & Hershenson, M.: The Psychology of Visual Perception. Holt, Rinehart & Winston, Oxford (1973)
6. Jain, A. K., Nandakumar, K., & Ross, A: Score Normalization in Multimodal Biometric Systems. 38(12), 2270—2285 (2005)

7. Jiang, B., Valstar, M., & Pantic, M: Action Unit Detection Using Sparse Appearance Descriptors in Space-time Video Volumes. Automatic Face and Gesture Recognition, Proc. IEEE Intl. Conf. on. (2011)
8. Liu, C., Yuen, J., & Torralba, A: SIFT Flow: Dense Correspondence across Scenes and Its Applications. Pattern Analysis and Machine Intelligence, IEEE Trans. on., 33(5), 978—994 (2011)
9. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I.: The Extended Cohn-Kande Dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression. Human Communicative Behavior Analysis, Workshop of CVPR. (2010)
10. Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J: Coding Facial Expressions with Gabor Wavelets. Automatic Face and Gesture Recognition, Proc. IEEE Intl. Conf. on. (1998)
11. Mckeown, G., Valstar, M. F., Cowie, R., & Pantic, M.: The Semaine Corpus of Emotionally Coloured Character Interactions. Multimedia and Expo, Proc. of IEEE Intl. Conf. on. (2010)
12. Ojansivu, V., & Heikkila, J.: Blur Insensitive Texture Classification Using Local Phase Quantization. Image and Signal Processing, Proc. of Int. Conf. on. (2008)
13. Pantic, M., & Rothkrantz, L: Automatic analysis of facial expressions: the state of the art. Pattern Analysis and Machine Intelligence, IEEE Trans. on., 22(12), 1424—1445 (2000)
14. Samal, A., & Iyengar, P. A.: Automatic recognition and analysis of human faces and facial expressions: a survey. Pattern Recognition. 22(1), (1992)
15. Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., & Pantic, M.: AVEC 2011 – The First International Audio/Visual Emotion Challenge. First International Audio/Visual Emotion Challenge and Workshop. Springer LNCS. (2011)
16. Snelick, R., Uludag, U., Mink, A., Indovina, M., & Jain, A. K.: Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems. Pattern Analysis and Machine Intelligence, IEEE Trans. on. 27(3), 450—455 (2005)
17. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., et al.: A comparative study of energy minimization methods for markov random fields with smoothness-based priors. Pattern Analysis and Machine Intelligence, IEEE Trans. on.. 30(6), 1068—1080 (2008)
18. Valstar, M. F., Jiang, B., Mehu, M., Pantic, M., & Scherer, K.: The First Facial Expression Recognition and Analysis Challenge. Face and Gesture Recognition, IEEE Int'l. Conf. (2011)
19. Valstar, M., & Pantic, M.: Induced disgust, happiness and surprise: an addition to the mmi facial expression database. Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect. (2010)
20. Viola, P., & Jones, M.: Robust Real-Time Face Detection. International Journal of Computer Vision. Intl. Conference on Computer Vision. (2001)
21. Yang, S., & Bhanu, B:. Facial expression recognition using emotion avatar image. The First Facial Expression Recognition and Analysis Challenge. Face and Gesture Recognition, IEEE Int'l. Conf. (2011)
22. Yu, J., & Bhanu, B.: Evolutionary feature synthesis for facial expression recognition. Pattern Recognition Letters 27, (2006)
23. Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. 31(1), 39—58 (2009)