Facial Emotion Recognition With Expression Energy

Albert Cruz
Center for Research in Intelligent Systems
216 Winston Chung Hall
Riverside, CA, 92521-0425, USA
acruz006@student.ucr.edu

Bir Bhanu
Center for Research in Intelligent Systems
216 Winston Chung Hall
Riverside, CA, 92521-0425, USA
bhanu@ee.ucr.edu

Ninad Thakoor
Center for Research in Intelligent Systems
216 Winston Chung Hall
Riverside, CA, 92521-0425, USA
ninadt@ee.ucr.edu

ABSTRACT

Facial emotion recognition, the inference of an emotion from apparent facial expressions, in unconstrained settings is a typical case where algorithms perform poorly. A property of the AVEC2012 data set is that individuals in testing data are not encountered in training data. In these situations, conventional approaches suffer because models developed from training data cannot properly discriminate unforeseen testing samples. Additional information beyond the feature vectors is required for successful detection of emotions. We propose two similarity metrics that address the problems of a conventional approach: neutral similarity, measuring the intensity of an expression; and temporal similarity, measuring changes in an expression over time. These similarities are taken to be the energy of facial expressions, measured with a SIFT-based warping process. Our method improves correlation by 35.5% over the baseline approach on the frame-level sub-challenge.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Experimentation, performance, theory

Keywords

Computer vision, image representation, video analysis

1. INTRODUCTION

Facial emotion recognition has applications in medicine [6], video games, human-machine and computer-interaction and affective computing [4].

Challenge data sets, such as the 2nd International Audio/Visual Emotion Challenge and Workshop (AVEC2012) [11], have advanced state-of-the-art by standardizing the data used by algorithms. This grand challenge provides a common ground with which these algorithms can be compared. However, an algorithm has yet to be revealed that can correctly identify emotions when there is not sufficient training data for an individual, or that individual was not encountered in the training data. This was demonstrated in Maronidis et al. [9] with inter-database experiments, and such is the case in AVEC2012. A conventional approach that classifies the samples solely on their feature vector performs poorly. The decision surface has been trained on particular expressions from the individuals in training data. If a new expression is queried, such as if a previously un-encountered individual expresses an emotion in a different way, it would not be properly labeled by the model. This is a similar situation to the face recognition challenge in the Labeled Faces in the Wild data set, where most individuals have only a single sample in the data. Wolf et al. [12] proposed the use of similarities that captured additional information based on relationships with other samples. That approach and the approach detailed here fall into a specific category of learning titled, "learning with side information" [12], where a discriminating relationship between the samples is learned. We propose the following similarities to improve classification for facial emotion recognition:

Neutral Similarity (Sec. 2.2). This similarity addresses the question, "how intense is the emotion?" When an expression has not been encountered, it may be useful to measure the difference between the current frame and some reference of a face which is not expressive. This similarity is important in measuring the degree of an emotion, which is a new requirement with AVEC2012.

Temporal Similarity (Sec. 2.3). This similarity addresses the question, "how is the expression changing?" If there is a significant difference between the current frame and the previous frame, it may indicate particular emotions. E.g., a neutral expression followed by furrowing of the brows may indicate anger. This similarity also adds temporal information to the approach.

These proposed similarities are inspired by One-Shot and Two-Shot similarities [12], but are substantially different. One-Shot and Two-Shot similarities are both based on unlabeled background data, not deviation from a neutral face or temporal changes. Additionally, we compute these similarities with a SIFT-based warping process, whereas One-Shot and Two-Shots are computed with a vector difference.

The rest of the paper is organized as follows: Sec. 2 details the approach for measuring expression energy, and Sec. 2.5

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.

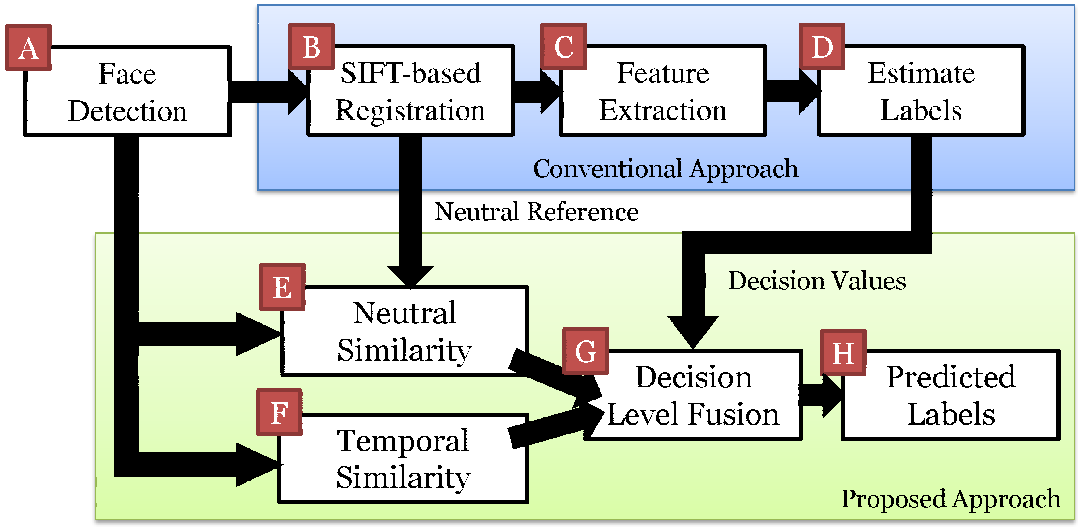

Figure 1: System overview.

details the appearance features also used in the approach. Sec. 3 details the experimental parameters used in the paper and presents the results on AVEC2012. A discussion is given in Sec. 4 that compares the appearance features and also compares the feature space of a conventional approach and the proposed approach. A conclusion is given in Sec. 5.

1.1 Contribution

Our contribution is the following: (1) two novel similarities, neutral and temporal similarity, to be used in unconstrained scenarios where samples in testing were not encountered in training; (2) a novel SIFT-Flow energy function for quantification of these similarities; and (3) improvement of the average correlation of emotion detection on AVEC2012 for both the frame-level and word-level sub-challenges.

2. TECHNICAL APPROACH

The proposed system approach is given in Fig. 1: (A) frames are extracted from the video and face region of interest is detected on a per-frame basis with a cascade of Haar-like features. In a conventional approach, (B) faces would be registered. We register faces with Avatar Image Registration. After registration, (C) an ensemble of low-level appearance features are extracted, re-projected with PCA and then fused at the feature level and (D) a model is trained with Support Vector Regression to estimate labels. With the proposed approach, in a separate pipeline, (E) neutral similarity is measured from a neutral reference that has been estimated during registration, and (F) temporal similarity is measured between frames. (G) These two similarities and the decision values from the conventional approach are fused to calculate (H) the emotion labels.

2.1 Facial Extraction and Registration

Face region of interest (ROI) is detected with Viola and Jones. Within each video, ROI detection errors are pruned with the following schema: a mean face image is generated for that video. If a frame in that video has an L_2 difference from the mean face image of more than four standard deviations away from the mean difference of the mean face image, it is considered an ROI detection error. In training, the frame is removed when training a model. In testing, the frame is not classified; its emotion label is taken to be the nearest frame with a successfully extracted ROI. Four standard deviations was determined empirically, to be robust versus face morphology and slight translation errors from Viola and Jones while removing ROI's that were not a whole face.

The extracted face ROI is aligned with Avatar Image Registration (AIR). In AIR, faces are aligned with SIFT-Flow [7] to an estimated neutral face, titled the Avatar Reference Image. The reader is referred to Cruz et al. [2] for a more in depth explanation of AIR.

2.2 Neutral Similarity

Neutral similarity measures the intensity of an emotion. In the conditions described in the motivation in Sec. 1, the difference of the current frame from a neutral face may provide discriminative information. The metric should indicate when facial expressions are close to neutral, such as in Fig. 2-B, and when facial expressions are intense, such as in C. This is computed by comparing a given frame, $I(\mathbf{x}, t)$, to some reference of a neutral face, such as A. The Avatar Reference Image, a neutrally expressive reference image, is taken to be the neutral reference. It is the mean face image generated from SIFT-registered face ROI's across all training videos and individuals. The reader is referred to Cruz et al. [2] for a proof of this as an estimate of a neutrally expressive face. The similarity is measured with the expression energy function to be described in Sec. 2.4.

2.3 Temporal Similarity

Temporal similarity measures the amount of change between frames. It provides additional information of how expressions are changing temporally. In Fig. 3 there are two pairs of frames taken from the sam video at different times, both spaced 1s apart. Between A and B, the individual is smirking, and the metric should detect the small change. Between C and D, there are many changes in expression, and the metric should reflect this with a higher energy than the energy given to frames A and B. This simi-

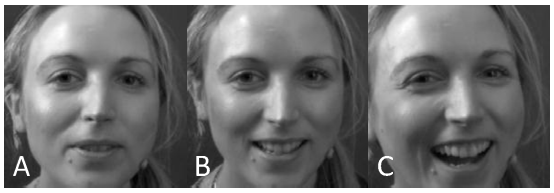

Figure 2: (From left to right): (A) A neutral face. (B) A slightly expressive face. (C) A very expressive face. The energy for C should be higher than B.

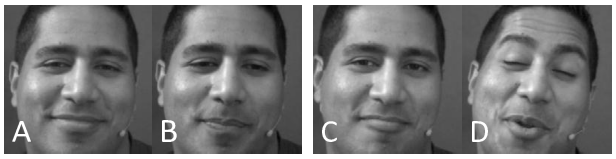

Figure 3: Frames spaced 1s apart. (A-B) A face with slight change in expression. (C-D) A face with large change in expression. The energy of A to B should be lower than for C to D.

larity is computed by comparing $I(\mathbf{x}, t)$ with $I(\mathbf{x}, t-\delta)$. It is also measured with the expression energy function.

2.4 Measuring Expression Energy

Let I_1 and I_2 be two images. We want to quantitatively measure the similarity between two expressive face ROI's. We propose measuring the similarity as the energy of the cost function of a SIFT-based warping process. The energy function used in the warping process can be used as a metric to take this measurement. This similarity is more meaningful than a distance metric between two feature vectors because SIFT-Flow warps the intensity values to match the two images, based on SIFT features. The original SIFT-Flow energy function [7] is as follows:

$$\begin{aligned} E(\mathbf{w}) = & \sum_{\mathbf{x}} \min \| s_1(\mathbf{x}) - s_2(\mathbf{x}+\mathbf{w}(\mathbf{x})) \|_1 \\ & + \gamma \sum_{\mathbf{x}} \left((u(\mathbf{x}))^2 + (v(\mathbf{x}))^2 \right) \\ & + \alpha \sum_{(\mathbf{x},\mathbf{y}) \epsilon N_{\mathbf{x}}} \left(\min |u(\mathbf{x}) - v(\mathbf{y})| \right. \\ & \left. + \min |u(\mathbf{x}) - v(\mathbf{y})| \right) \end{aligned} \quad (1)$$

where $\mathbf{x}$ is a pixel in the image; $\mathbf{w}(\mathbf{x})$ is the motion vector at pixel $\mathbf{x}$ between I_1 and I_2; $\mathbf{w}(\mathbf{x}) = |u(\mathbf{x}), v(\mathbf{x})|$; and s_1 and s_2 are the dense SIFT descriptors of I_1 and I_2 respectively; $\mathbf{y}$ is a pixel that iterates for each member of the $N_{\mathbf{x}}$ neighborhood of $\mathbf{x}$; and γ is a normalization constant. The first summation term measures the L_1 difference between the two SIFT descriptors. The second summation term is the squared magnitude of the motion vector and measures the magnitude of motion in between the two images. The third summation term enforces homogeneity. We propose the following, modified SIFT-Flow energy function:

$$\begin{aligned} E(\mathbf{w}) = & \sum_{\mathbf{x}} \min \| s_1(\mathbf{x}) - s_2(\mathbf{x}+\mathbf{w}(\mathbf{x})) \|_1 \\ & + \gamma \sum_{\mathbf{x}} \left((u(\mathbf{x}) - \mu_u)^2 + (v(\mathbf{x}) - \mu_v)^2 \right) \end{aligned} \quad (2)$$

where μ_u and μ_v are the mean of $u(\mathbf{x})$ and $v(\mathbf{x})$ respectively. Eq. 2 differs from the original SIFT-Flow energy function in the following ways: (1) the second term has the mean vector subtracted from u and v. Viola and Jones is accurate but not precise; the face ROI suffers from translation errors. The mean is subtracted from $\mathbf{w}$ to reduce the impact of a whole-plane translation from the ROI translation errors. (2) There is no term that measures the homogeneity of motion. The homogeneity term in the original implementation of SIFT-Flow is enforced to improve visual quality of the warping result; we seek the similarity between two images and do not care for visualization of the result.

An example of neutral expression energy is given in Fig. 5. The frames are ordered according to neutral similarity, measured with expression energy, from left to right, top to bottom. The faces in the first row have close to neutral expression, whereas the faces in the bottom row have strong expressions. An example of temporal expression energy is given in Fig. 4. Temporal expression energy correctly indicates the amount of change between frames: in Fig. 4A, there is only mouth movement, whereas in I, there is movement of pose, gaze, cheeks, etc.

The energy for neutral and temporal similarities and the decision values from a conventional approach are fused at the decision level with support vector regression (SVR). Temporal smoothing is introduced by aggregating all the similarities in a window: at t, the three similarities about t in a window size of l are concatenated to form the features for the SVR.

2.5 Conventional Approach

We employ four appearance features: (1) Uniform Local Binary Patterns (ULBP). An image is encoded in terms of micro-patterns at the pixel-neighborhood level [1]. (2, 3) Three-Patch (TPLBP) and Four-Patch Local Binary Patterns (FPLBP). These features were proposed by Wolf et al. [12] as features that were complimentary to LBP features. In TPLBP, a pixel's neighborhood is compared to neighborhoods about the pixel. In FPLBP, the pixel's neighborhoods are compared to further neighbors. (4) Discrete Cosine Transform (DCT) [8], the DCT of the face. It was shown that a feature fusion of LBP, TPLBP and FPLBP improved classification rates on the non-trivial Labeled Face in the Wild data set [12]. For that reason, we hypothesize that this fusion captures meaningful facial appearance information, and we employ a similar fusion for facial emotion recognition. We re-project the feature vectors for each feature set with PCA before concatenating them to reduce dimensionality.

2.5.1 Local Binary Patterns

Local Binary Patterns (LBP) are a dense low-level appearance feature that have been used for both face recognition and facial emotion recognition [1]. While the method used to generate the features is commonly referred to as LBP, the features are actually a histogram of an LBP image. In classical 8-neighborhood LBP images, the pixel value $\mathbf{x}$ of a

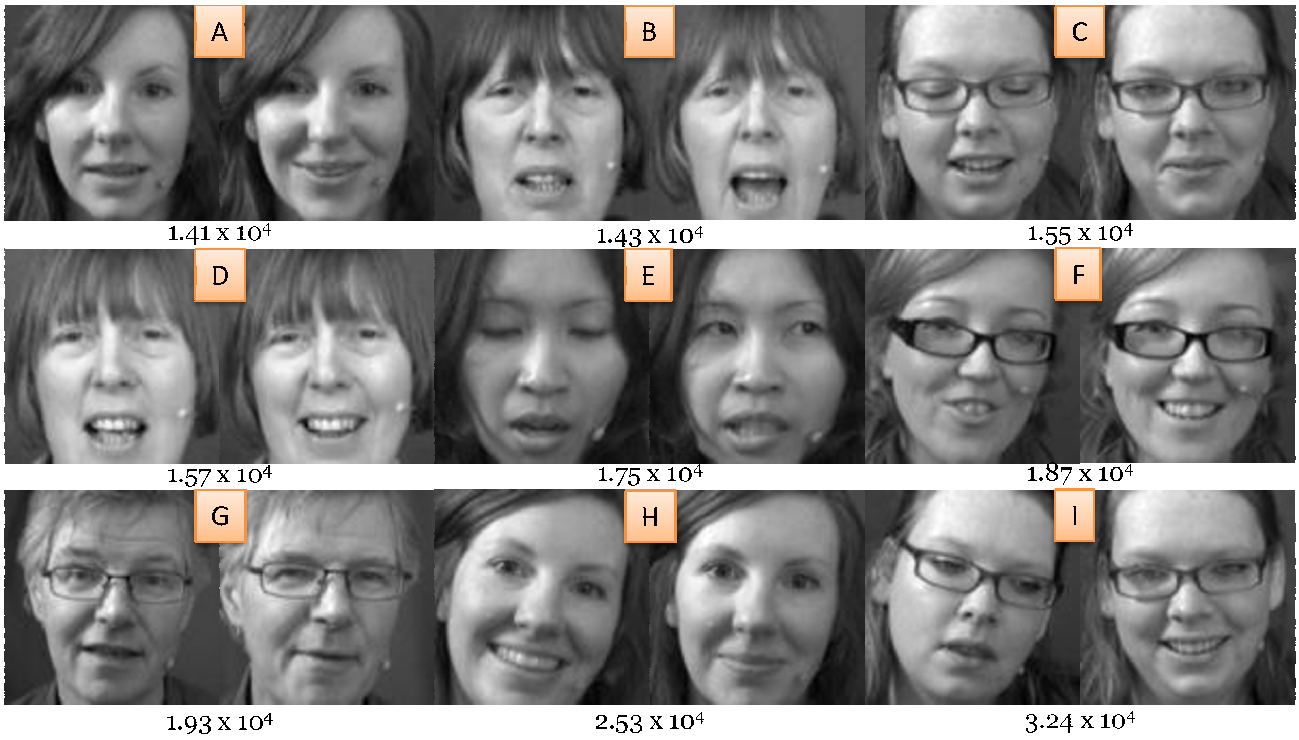

Figure 4: An example of temporal similarity measured with expression energy. Images are sorted from left to right, then top to bottom in order of neutral to most expressive. Beneath the frame is the measured expression energy.

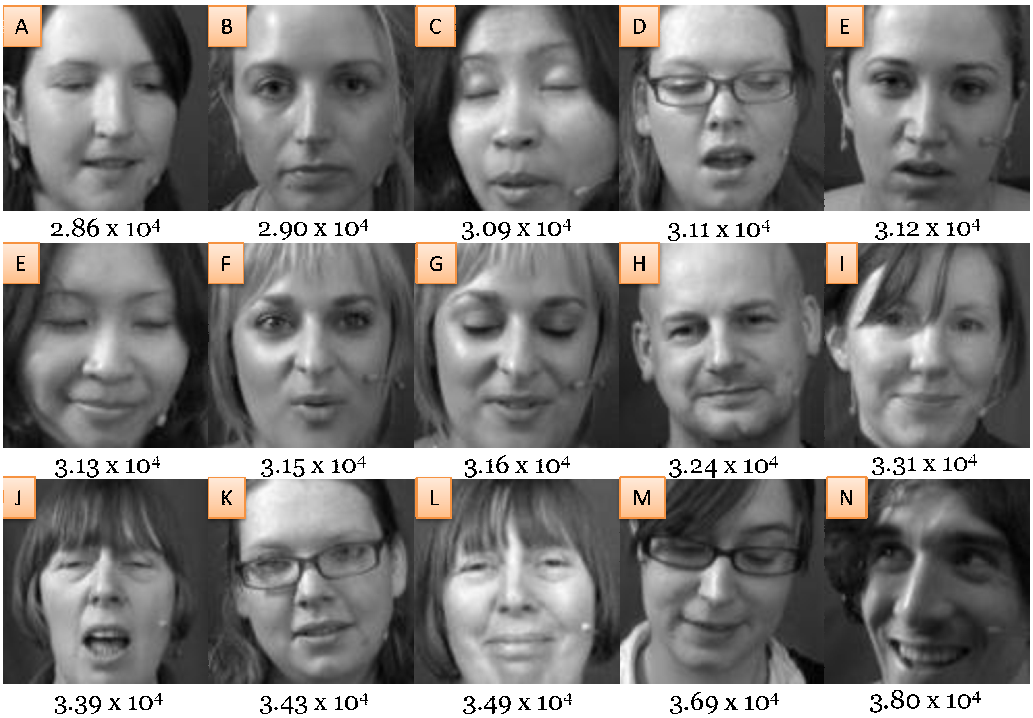

Figure 5: An example of neutral similarity measured with expression energy. Images are sorted from left to right, then top to bottom in order of neutral to most expressive. Beneath the frame is the measured expression energy.

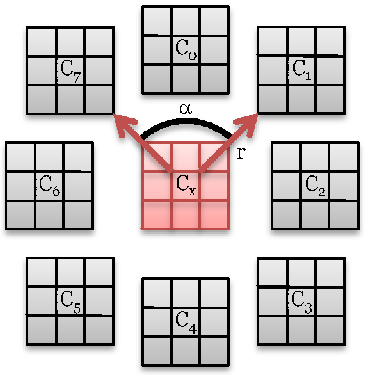

Figure 6: An example of TPLBP.

grayscale image is compared to its 8 member neighborhood $N_{\mathbf{x}}$. It is characterized by three steps: (1) Let $\texttt{LBP}(\mathbf{x})$ be the LBP image. It takes on an 8 bit integer, where each bit reflects whether or not the intensity value at $\mathbf{x}$ is greater than one of its neighbors:

$$\texttt{LBP}(\mathbf{x}) = \sum_{i=1}^{8} 2^{i-1} H\left(\langle N_{\mathbf{x}} \rangle_i - I(\mathbf{x})\right) \quad (3)$$

where $H(.)$ is a step function and $\langle N_{\mathbf{x}} \rangle_i$ is the i-th neighbor of $\mathbf{x}$. (2) Assuming an 8-bit gray level image, a 256-bin histogram is generated from $\texttt{LBP}(\mathbf{x})$. It was found that a simplification to 59-bins results in a rotation invariant property [10], and this simplification is referred to as Uniform Local Binary Patterns (ULBP). (3) To account for face morphology, a face image is reduced to $M \times M$ evenly sized local regions.

2.5.2 Three-Patch LBP

Whereas ULBP encodes micropatterns–how a pixel $\mathbf{x}$ compares to its immediately surrounding patch of pixel values C, TPLBP encodes how a the neighboring patches of C are changing with each other, i.e. how homogenous the region about C is in terms of texture. That is, ULBP encodes a micropattern in a compact 8-bit representation and TPLBP encodes a larger pattern and homogeneity, also with a compact 8-bit representation. To compute TPLBP, the values of three patches are compared to produce a single bit value in the code assigned to $\mathbf{x}$. Let $C_{\mathbf{x}}$ be an 8-neighborhood patch of a given pixel. S is the number of patches surrounding $C_{\mathbf{x}}$. The patches are evenly distributed about $\mathbf{x}$ along a circle of radius r. The encoding compares two patches spaced α apart and compares their values to $C_{\mathbf{x}}$. This is done S times, resulting in S bits per pixel. A visual example is given in 6.

$$\texttt{TPLBP}(\mathbf{x}) = \sum_{i=1}^{S} 2^{i-1} H(d(C_i, C_{\mathbf{x}}) - d(C_{i+\alpha \bmod S}, C_{\mathbf{x}}) - \tau) \quad (4)$$

where C_i and $C_{i+\alpha \bmod S}$ are two patches along the ring; $C_{\mathbf{x}}$ is the central patch; the function $d(.)$ is the L_1 distance function between two patches; $H(.)$ is the step function; and τ is an offset to make Eq. 4 robust to small intensity changes due to noise.

2.5.3 Four-Patch LBP

FPLBP continues the comparison of texture about $C_{\mathbf{x}}$ by further comparing the neighbors of $C_{\mathbf{x}}$ to their neighbors. It resembles TPLBP, except S patches are examined of a radius r_1 from $\mathbf{x}$, and an additional S patches examined of a radius r_2 from $\mathbf{x}$. Instead of comparing two pairs of patches from the same ring, a pair consists of a patch in ring 1 and a patch in ring 2. The second pair is circularly symmetric to the first pair. A bit is encoded based on which pair has a greater difference. An visual example is given in Fig. 7. The formal definition of the FPLBP code is as follows:

$$\texttt{FPLBP}(\mathbf{p}) = \sum_{i=1}^{S} 2^{i-1} H(d(C_{1,i}, C_{2,i+\alpha \bmod S}) - d\left(C_{1,i+S/2}, C_{2,i+S/2+\alpha \bmod S}\right) - \tau) \quad (5)$$

where the variables are the same as in Eq. 4; except $C_{i,j}$ is the ith patch surrounding $C_{\mathbf{x}}$, r_j away from $C_{\mathbf{x}}$.

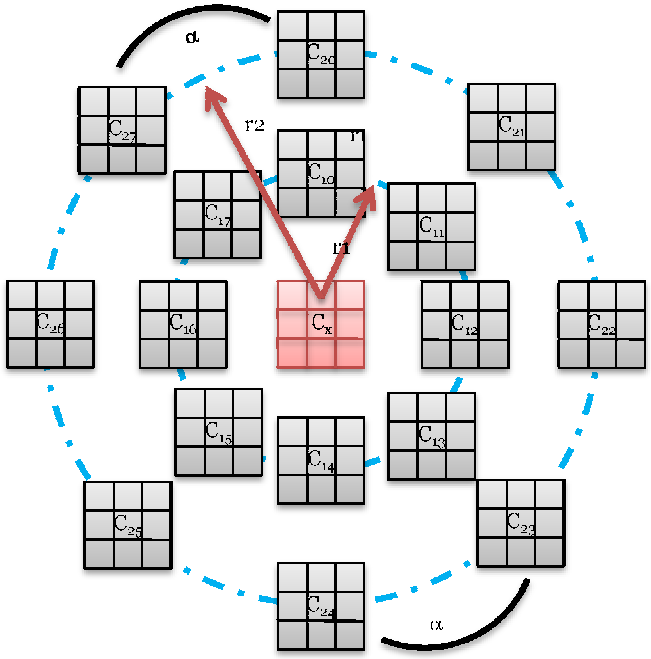

Figure 7: An example of FPLBP.

3. EXPERIMENTATION

The reader is referred to the challenge paper for a description of the data [11]. This is a video only approach that computes features at the frame level. Videos are sub-sampled at 1fps. For training, we use a two-fold cross-validation, where the training set is used to classify the development set, and vice versa. The reported correlation scores are the mean and standard deviation of the correlation of both folds: $E\langle\rho\rangle \pm \text{Std}\langle\rho\rangle$. Avatar Image Registration is run for three iterations, which is empirically selected from previous work [2]. LBP is computed at a radius of 1 and for 8 neighbors, with a 59 bin simplificiation s.t. the feature is rotationally invariant. For TPLBP, $\alpha = 2$, $S = 8$ and $w = 3$. For FPLBP, $\alpha = 2$, $S = 8$ and $w = 3$. For TPLBP and FPLBP, these parameters were empirically selected in previous work. For PCA, we rank the eigenvectors by eigenvalue and retain

Table 1: Training Results on AVEC2012 for Feature Fusion

	Arousal	Expectancy	Power	Valence	Avg.					
	With Extreme Learning Machine					With Support Vector Machine				
LBP	.221 ± .096	.027 ± .007	.037 ± .044	.136 ± .030	.105	**.380 ± .006**	.166 ± .007	.031 ± .014	**.263 ± .122**	.210
TPLBP	.064 ± .063	.021 ± .020	.072 ± .067	.058 ± .052	.054	.292 ± .051	.164 ± .019	.046 ± .060	.198 ± .067	.183
FPLBP	.139 ± .098	.022 ± .098	.026 ± .027	.051 ± .010	.059	.311 ± .046	.165 ± .042	**.094 ± .077**	.237 ± .014	.198
DCT	.221 ± .026	.019 ± .007	.007 ± .001	.058 ± .006	.076	.258 ± .026	.170 ± .007	.017 ± .022	.183 ± .103	.163
Fusion	.133 ± .016	.045 ± .011	.022 ± .031	.059 ± .031	.065	.323 ± .034	**.189 ± .027**	.046 ± .012	.219 ± .005	.194

Table 2: Training Results for Proposed Similarities

	Arousal	Expectancy	Power	Valence	Avg.
Video-Only Baseline					
Conventional	.380 ± .160	.189 ± .049	.094 ± .077	.263 ± .126	.231
Proposed	.454 ± .042	.176 ± .043	.139 ± .064	.378 ± .099	.264

Table 3: Official AVEC2012 Testing Results

	Ar.	Exp.	Pow.	Val.	Avg.
Video-Only Baseline					
FCSC	.092	.121	.064	.140	.104
WLSC	.091	.114	.121	.143	.117
Proposed Approach					
FCSC	.227	.093	.102	.141	.141
WLSC	.258	.049	.039	.153	.125

the eigenvectors that account for 95% of the total variance. For similarity fusion, $\gamma = .008$ and $l = 5$.

3.1 Results

For the conventional method, SVR with a Radial Basis Function kernel and Extreme Learning Machine [5]–a supervised learning algorithm based on a single hidden layer feedforward network–are compared in Tab. 1. Tab. 1 also compares which features are suitable for each class. In Tab. 2, we compare the performance of the best performing features when using a conventional approach versus the proposed similarity fusion. For Arousal and Valence we use LBP features; Expectancy, feature fusion; and Power, FPLBP. All classes employ SVR. The official testing results are given in Tab. 3, where FCSC indicates frame-level scoring and WLSC indicated word-level scoring.

4. DISCUSSION

From Tab. 1, SVR with a radial basis function kernel is the best performer for all classes and features. We postulate that the feature space was not discriminative, and had to be projected into a more highly dimensional space with SVR's radial basis function. Also, from Tab. 1, feature fusion of all the features sets does not always increase performance. Feature fusion of LBP, TPLBP, FPLBP and DCT features gives better classification performance and the least variance for Expectancy. However, for Arousal and Valence, LBP is the best performer. For expectancy, FPLBP is the best classifier. We use best performing feature for each class. FPLBP's performance is surprising because Wolf et al. [12] designed FPLBP to be a complimentary feature to be fused with LBP and TPLBP–their experimental results showed no performance gain when using FPLBP alone for a face recognition task. FPLBP encodes large edges. In Fig. 8, large facial features such as folds in the cheek, eyebrows and lips are encoded. In the Fontaine emotion model [3], anger and sadness are on opposite ends of the spectrum. These two emotions elicit expressions that strongly distort the large facial feature on the face such as the mouth and eyebrows, e.g., in anger the eyebrows are furrowed and the mouth is open with teeth clenched. We posit that FPLBP better encodes these expressions versus other features, leading to the increase in performance.

Figure 8: FPLBP encodes large edges.

For the official testing results in Tab. 3, the proposed approach doubles the correlation with arousal and power emotional states. Neutral and temporal similarities provide a more discriminative feature space than feature fusion alone. Note that neutral and temporal similarities are computed pre-registration. In Fig. 9, we provide examples comparing the feature space of a conventional method with feature fusion and the feature space of the proposed neutral and temporal similarities for arousal, expectancy and valence. The conventional feature space has the feature corresponding to the highest eigenvalue along the y-axis and the second highest eigenvalue along the x-axis. The proposed similarity feature space has neutral similarity along the x-axis and temporal similarity along the y-axis. The color indicates a quantized emotional state into bins of high and low, similar to AVEC2011–high is greater than the mean, low is less than the mean value of the label. Note that for all classes the conventional feature space is not very discriminative. It would require a complex decision surface. This is not so

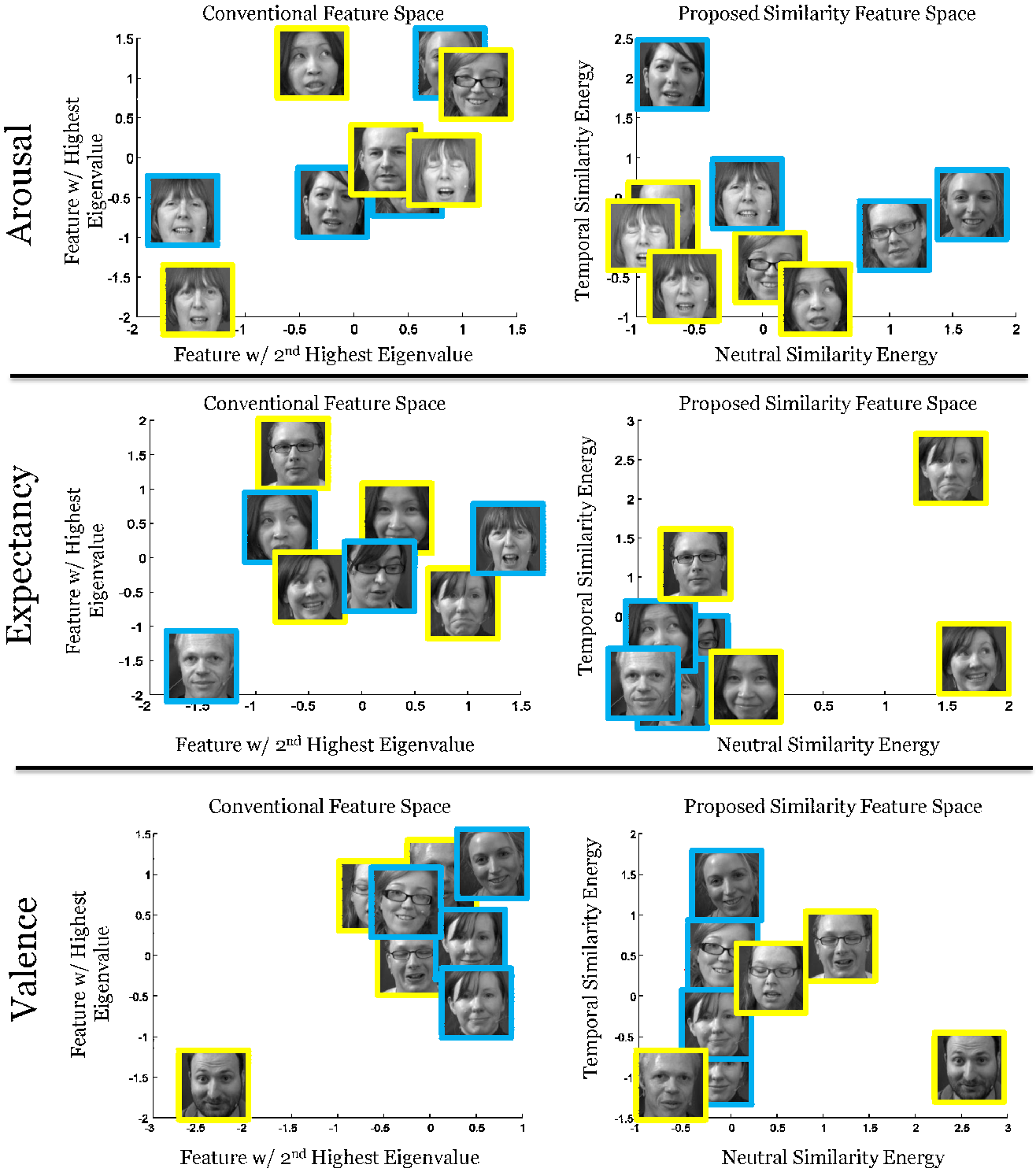

Figure 9: (Left Column) Feature space of feature fusion with the conventional approach and (Right Column) feature space of the propose neutral and temporal similarities. Blue indicates a high value in that emotion, whereas and yellow indicates a low value.

with the proposed methods feature space, the differences of which are described below. Note that in each figure the axes were z-normalized.

For arousal, positive (blue) indicates love and anger, whereas negative (yellow) indicates contentment and disappointment. The proposed method creates a perfectly linearly separable feature space. Note that negative arousal tends to have lower values in neutral and temporal similarities. Intuitively, a person that is content or disappointed will not make as many expressions as a person who is expressing love or anger, two emotions that elicit strong expressions. This is visible in the feature space, as face frames with less than average arousal form a distinctly close to neutral, lacking motion area in the space.

For expectancy, positive (blue) indicates surprise, whereas negative (yellow) indicates guilt and other emotions. It should be noted that while surprise has a high value along this dimension–a value of 3– the majority of big-six emotional states fall densely packed near neutral, zero [3]. This is manifested by in the proposed methods feature space by a very dense concentration of positive expectancy exhibiting less motion and more closeness to neutral than in the previous case with arousal. We can infer from this figure that AVEC2012 contains mostly examples of subtle surprise, e.g., an individual is being mildly impressed or unsure of how to answer the embodied agent. The individual is not shocked; strong surprise would have a high value along the neutral similarity axis. Positive expectancy is located in all other parts of the feature space, corroborating that this emotion is best described as surprise versus other, as defined in the model [3].

For valence, positive (blue) indicates happiness and love, whereas negative (yellow) indicates contempt and disgust. The figure contains examples of happiness ranging from subtle happiness to over happiness. This class is an example of the quality of motion estimation from temporal similarity. The degree of happiness tends to be described by temporal change. A large smile means a large change, as the mouth is moving and it is a large facial feature.

5. CONCLUSION

We proposed novel similarities for facial emotion recognition to be used in scenarios where new individuals are introduced in the testing data. The expression energy measures these similarities with a modified SIFT-Flow energy function. We found that in conventional methods, feature fusion does not always lead to performance increases with AVEC2012. We found that the feature space created by neutral and temporal similarities bears a resemblance to the dimensions defined by the Fontaine emotion model. The proposed method increases correlation over the video-baseline by a ratio of 1.355.

6. ACKNOWLEDGEMENTS

Support for this work was provided for in part by NSF grants 0727129 and NSF IGERT: Video Bioinformatics Grant DGE 0903667. The contents and information do not reflect the position or policy of the U.S. Government.

7. REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] A. Cruz, B. Bhanu, and S. Yang. A psychologically-inspired match-score fusion model for video-based facial expression recognition. In *Affective Computing and Intelligent Interaction*, Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011.

[3] 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.

[4] H. Gunes and M. Piccardi. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems Science and Cybernetics Part B*, 31(1):64–84, 2009.

[5] G. B. Huang, H. Zhou, X. Ding, and R. Zhang. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems Science and Cybernetics Part B*, 42(2):513–529, 2012.

[6] R. E. Kaliouby and P. Robinson. The emotional hearing aid: an assistive tool for children with asperger syndrome. *Universal Access in the Information Society*, 4(2):121–134, 2005.

[7] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011.

[8] L. Ma and K. Khorasani. Facial expression recognition using constructive feedforward neural networks. *IEEE Transactions on Systems Science and Cybernetics Part B*, 34(3):1588–1595, 2004.

[9] A. Maronidis, D. Bolis, A. Tefas, and I. Pitas. Improving subspace learning for facial expression recognition using person dependent and geometrically enriched training sets. *Neural Networks*, 24(8):814–823, 2011.

[10] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[11] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic. Avec 2012 Ű the continuous audio/visual emotion challenge. In *Proc. Second International Audio/Visual Emotion Challenge and Workshop (AVEC 2012), Grand Challenge and Satellite of ACM ICMI 2012*, 2012.

[12] L. Wolf, T. Hassner, and Y. Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1978–1990, 2011.