Sparse Temporal Representations for Facial Expression Recognition

S.W. Chew[1], R. Rana[1,3], P. Lucey[2], S. Lucey[3], and S. Sridharan[1]

[1] Speech Audio Image and Video Technology Laboratory at Queensland University of Technology, Australia
[2] Disney Research Pittsburgh
[3] Commonwealth Science and Industrial Research Organisation (CSIRO), Australia
{sien.chew,s.sridharan}@qut.edu.au, patrick.lucey@disneyresearch.com,
{rajib.rana,simon.lucey}@csiro.au

Abstract. In automatic facial expression recognition, an increasing number of techniques had been proposed for in the literature that exploits the temporal nature of facial expressions. As all facial expressions are known to evolve over time, it is crucially important for a classifier to be capable of modelling their dynamics. We establish that the method of sparse representation (SR) classifiers proves to be a suitable candidate for this purpose, and subsequently propose a framework for expression dynamics to be efficiently incorporated into its current formulation. We additionally show that for the SR method to be applied effectively, then a certain threshold on image dimensionality must be enforced (unlike in facial recognition problems). Thirdly, we determined that recognition rates may be significantly influenced by the size of the projection matrix $\mathbf{\Phi}$. To demonstrate these, a battery of experiments had been conducted on the CK+ dataset for the recognition of the seven prototypic expressions − anger, contempt, disgust, fear, happiness, sadness and surprise − and comparisons have been made between the proposed temporal-SR against the static-SR framework and state-of-the-art support vector machine.

Keywords: sparse representation classification, facial expression recognition, temporal framework

1 Introduction

Advancements made in the field of affective computing research are being rapidly propelled by commercial interests such as marketing, human-computer-interaction, health-care, security, behavioral science, driver safety, etc. A central aim of this research is to enable a computer system to detect the emotional state of a person through various modalities (e.g., face, voice, body, actions), in which the inference through one's facial expression had been a significant contribution. In the recent literature [1,2,3,4], an increasing number of machine learning techniques had been proposed to take advantage of the dynamics inherent in facial expressions. Intuitively, enabling the temporal information of a signal to be exploited serves to elegantly unify a machine learning framework

Y.-S. Ho (Ed.): PSIVT 2011, Part II, LNCS 7088, pp. 311–322, 2011.

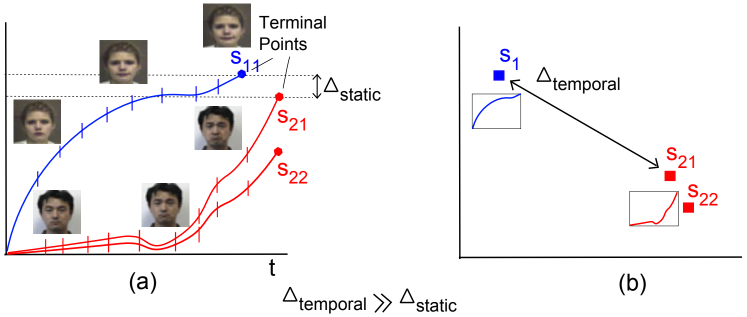


Fig. 1. As a simple thought experiment, consider in (a) a static approach which distinguishes between 2 classes of signals s_1 and s_2 (e.g., happiness vs sadness as the choice of signals here). Observe that only the terminal points in (a) are considered, where the signal's temporal evolution have been discarded (hence small Δ_{static}). On the other hand, a strategy which exploits temporal information is able to map (a) to (b), which enforces both signal classes to occupy a two-dimensional space based on their 'shape' in time (i.e., temporal content); and thus produces a large Δ_{temporal}.

with the architecture of how facial expressions naturally evolve (see Figure 2). To illustrate this concept, observe in Figure 1(a) how the training/prediction of a classifier is determined using only the terminal points (e.g., a single frame containing the expression's apex). A problem with such a static approach is that it takes into account only a single state of the signal, but disregards the signal's past states (i.e., memoryless). By incorporating temporal information (Figure 1(b)), one is able to amplify the minuscule static differences Δ_{static} using the signals' temporal content to obtain a larger Δ_{temporal}. A drawback with adopting such a strategy revolving round a temporal framework, however, is that the complexity of the problem increases proportionally to the quantity of temporal information under consideration (i.e., more training and testing data). Having this in mind, we aim to develop a method which achieves these objectives, while at the same time being able to reduce data dimensionality. One method which gracefully fulfills the latter requirement is the method of sparse representation (SR) classification [5]. However, this method in its current formulation is unable to fulfill the former objective of modelling the dynamics of various expressions. In this paper, we propose a temporal framework for it to fully exploit the dynamics of facial expression signals in an efficient manner (see Section 3).

Recently, the above-mentioned (static) SR classification method had generated considerable excitement in the field of face recognition, thus its transition towards facial expression recognition comes as little surprise. In [5], it was proposed that the downsampling of the input image to a dimensionality of approximately 10×10 pixels, and then projecting this image using a random projection produced impressive performance for the task of person identification. From our experiments, we found that this procedure was not suitable for expression recognition. As opposed to solving for the identity of a person, facial expressions are formed through numerous interactions between various facial muscle groups

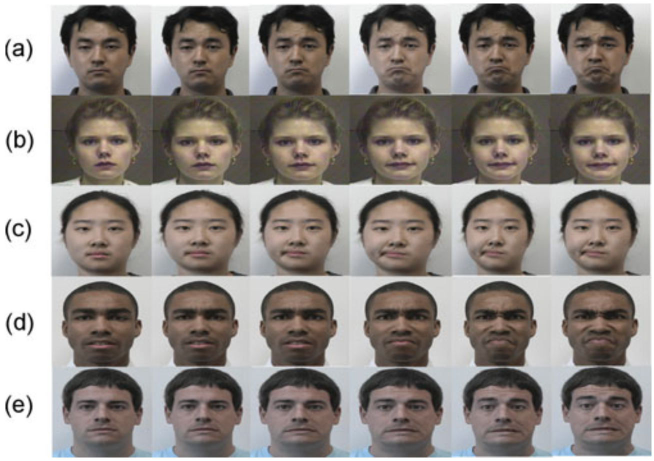


Fig. 2. Examples from the CK+ dataset [10] illustrating the strong temporal links present within neighbouring frames among different expressions, (a) sadness, (b) happiness, (c) contempt, (d) anger and (e) fear

(e.g., eyebrows, lips, nose), most of which require an adequate number of pixels to represent.

In Section 3 we show that the solution to such high dimensionality vectors is computationally exhaustive. We employ SR theory to reduce the image dimensionality, and investigate the impact of different dimensions on recognition performance. Interestingly, different expressions were observed to react differently to this. In most of the works [6,7,8,9] pertaining to using SR classifiers for expression detection, a static approach had been adopted; that is, only single independent frames from various sequences were used. However, it should be recognized that facial expressions are inherently temporal by nature (as shown in Figure 2) and it will be beneficial to incorporate temporal information into the SR classifier.

The central contributions of this paper are,

- Propose a temporal framework for sparse representation classifiers to improve facial expression recognition rates.
- Investigate the effects of downsampling the input images, and the significance of dimensionality reduction on detection accuracy.
- Compare the state-of-the-art SVM framework versus the conventional static-SR classifier and the proposed temporal-SR classifier frameworks and demonstrate that the proposed temporal method offers improved recognition rates. To the best of our knowledge, we are the first to quantitatively report the performance of SR classifiers for all seven expressions. This is important because application developers can decide between SVM and temporal-SR classifier based on the accuracy versus complexity trade-off.

2 Review of Temporal-Based Methods in Expression Recognition

Exploiting temporal information for facial expression recognition is not new. For example, in [4] which used a dynamic Bayesian network to model the dynamic evolution of various facial action units (AUs). A slightly different approach was adopted in [3] which modeled AUs using expression dynamics coupled with phase information. A list of other temporal techniques can be found in [11,12,13]. The major difference between these works and ours, however, is that all these methods proposed require complex features to be extracted from multiple temporal frames. As feature extraction may be considered to be computationally expensive even in the static context, the problem becomes additionally complex and computationally expensive when multiple temporal frames are to be considered. On the other hand, our method does not require any feature representations to be computed. Furthermore, our method is driven by SR theory which is different from all of the above-mentioned methods. SR classification had been used for both face recognition [5] and facial expression recognition [14] previously (especially in [5] which utilized random features). However, none of these SR methods had capitalized on expression dynamics. Our proposed method exploits expression dynamics through SR theory, and we demonstrate the advantages of this method over other feature-based alternatives that rely mainly on only spatial information.

3 A Temporal Framework for Sparse Representation Classification

In this section we describe the underlying mechanics of the proposed Sparse Representation (SR) classifier for facial expression recognition. We initiate the description in terms of static frames, and then introduce the incorporation of temporal information into the framework.

Let us consider that we have N facial expression images spanning over $c = 1, 2, \ldots, C$ class. We represent these N images by N vectors $\vec{v}_1, .., \vec{v}_N \in \mathbb{R}^n$. Let us construct a dictionary ξ by packing the vectors $\vec{v}_i, \forall_{i=1,..,N}$ into the columns of a matrix $\xi \in \mathbb{R}^{n \times N}$. Intuitively, a test sample (e.g., a face image) $\vec{\gamma} \in \mathbb{R}^n$ of class $i \in \{1, 2, ..., C\}$ can be represented in terms of the dictionary ξ by the following linear combination,

$$\vec{\gamma} = \xi \vec{\alpha}, \tag{1}$$

where $\vec{\alpha} = [0, 0, \pi_{i1}, .., \pi_{ik}, 0, 0]^T$, π_{ij} are some scalars and k is the number of face images per class. Clearly, the solution to (1) (i.e. $\vec{\alpha}$) would recognize the test image class (class corresponds to nonzero element is the match), however, we have to identify a method to compute a sparse $\vec{\alpha}$.

A general method to find the sparse solution of (1) is to solve the following optimization problem:

$$\arg\min \hat{\alpha} = ||\vec{\alpha}||_0, \quad s.t. \vec{\gamma} = \xi \vec{\alpha}, \tag{2}$$

where $||.||_0$ denotes the ℓ_0 norm, which returns the nonzero elements of α. Note that if $n > N$, the system is overdetermined, in that case (2) can be solved in polynomial time. Typically the dimension of the image (n) is quite high compared to the available image set, therefore, it is rather impossible for normal computers to solve (2) [5]. For this reason, in practice, the dimension of n is reduced to a smaller size $d << n$ (by multiplying a random projection matrix $\varPhi \in \mathbb{R}^{d\times n}$ with ξ), which turns (2) into a underdetermined problem. In general, searching for a sparse solution of an underdetermined system using (2) is NP-hard.

Encouragingly, SR theory shows that if $\vec{\alpha}$ is sufficiently sparse, then this underdetermined system can be solved using the following ℓ_1 norm minimization problem, which will produce a similar solution to solving the ℓ_0 norm.

$$\arg\min \vec{\alpha}^* = ||\vec{\alpha}||_1, \quad s.t. \vec{\gamma} = \xi\vec{\alpha} \tag{3}$$

However, a sparse $\vec{\alpha}$ cannot always guarantee a unique solution to (3). SR theory shows that if $\varTheta = \varPhi\xi$ obeys the restricted isometry property(RIP) [15], then the underdetermined system (1) can be solved through (3). More encouragingly, SR theory also suggests that $\varPhi$ obeys RIP. Such as, we use a $\varPhi$ which is populated by sampling normally distributed numbers with zero mean and variance $\frac{1}{d}$ (i.e., $\varPhi \sim \mathcal{N}(0, \frac{1}{d})$). SR theory has shown that $\varPhi \sim \mathcal{N}(0, \frac{1}{d})$ obeys RIP when $d \propto K \log(\frac{n}{K})$. Here K is the measure of the sparsity of α. In this paper, instead of seeking to determine an optimal d, we investigated the impact of different values of d on the detection accuracy (see Figure 5).

Transitioning to a temporal framework, the most straightforward approach to incorporate the dynamics of a video sequence (i.e., temporal frames) into a SR classifier would be to simply concatenate consecutive frames into ξ, such that $\widetilde{\xi} \in \mathbb{R}^{n\times(N\mapsto\tau)}$, where $\tau = Nt$ and t represents the length of the temporal window. Intuitively, one problem with this approach is that the sparsity of the solution $\vec{\alpha}^*$ is ultimately reduced (by a factor of t) due to the increase in dimensionality of ξ. Following this, γ is then said to be composed of an *additional number of terms* in the linear combination $\xi\vec{\alpha}$ (which may be considered to be noise). In order to circumvent this problem, we postulate that the dimensions of ξ must be maintained to be at N number of columns.

In order to retain the size to N, each column in $\widetilde{\xi}$ is formed through the fusion of multiple frames into a single frame that is representative of the temporal information in all t frames; such that $t \mapsto 1$ and $\widetilde{\xi} \in \mathbb{R}^{n\times\tau} \mapsto \widetilde{\xi} \in \mathbb{R}^{n\times N}$. Another point that needs to be addressed is whether the utilization of sparse feature representations (i.e., a sparse $\widetilde{\xi}$) would lead to a better dynamic model. We explored a technique described in [7] which utilized absolute difference images (i.e., $|I_{\text{apex}} - I_{\text{neut}}|$) as sparse feature representations for ξ and γ . Theoretically speaking, an argument may be made that the majority of holistic information in the face is effectively removed once the absolute difference operation is performed. This may be undesirable because facial expressions are essentially holistic in nature (i.e., anger or happiness, etc, occurs in the whole face and are therefore not limited to specific local facial regions). To show this, we incorporated absolute difference feature representations into $\widetilde{\xi}$, and compared the performance with using just pixels,

$$\mathbf{D}_j = \left[\left| I_{(\text{apex}-t)} - I_{\text{neutral}} \right|, \left| I_{(\text{apex}-t+1)} - I_{\text{neutral}} \right|, \ldots, \left| I_{\text{apex}} - I_{\text{neutral}} \right| \right]; \quad (4)$$
$$\text{where} \quad j \in \{1, \ldots, \tau\}$$

where the columns $\mathbf{D}_j \in \mathbb{R}^{n \times t}$ of $\widetilde{\xi} \in \mathbb{R}^{n \times \tau}$ are calculated using absolute difference image operations on the neutral frame with respect to frames 1 to t. Empirical results shown in Figure 4(b) suggested that utilizing these sparse absolute difference features had led to a deterioration in detection accuracy. More details can be found in Section 5.

Understanding this, we focused on using only pixels (i.e., no features) and analyzed various approaches of temporal fusion to form the columns of $\widetilde{\xi}$. Simply computing the global average of all t frames in the temporal window would satisfy this criterion. However, from preliminary experiments conducted, we found that one drawback with such an approach was that a *blurring* phenomenon was induced as a result of registration-error/pixel-misalignments that is inherent in all tracked faces. In order to minimize blurring, we employed local averages of several selected facial regions which we had deemed vital in expression recognition (i.e., eyes and mouth, see Figure 3). The pixels residing outside of these two regions were then filled by the remainder of pixels from the apex frame. Further details on this approach are discussed in Section 5.

More sophisticatedly, temporal fusion can be accommodated by familiar methods such as principal component analysis (PCA) and discrete wavelet transforms, etc. To elucidate, the edges in an image may be considered to be the most salient features [16] perceived by the human visual system. Wavelet decomposition may be employed to extract these salient features, and the combination of these features would thus effectively capture the dynamics of the entire window into a single frame. Similarly with PCA, the most salient features are computed using the eigenvalues of the respective covariance matrices. The complete list of all temporal fusion methods employed in this paper is listed as follows – a) local average (AVG), b) gradient pyramid (GRA), c) laplacian pyramid (LAP), d) principal component analysis (PCA), e) discrete wavelet transform (DWT), f) shift invariant discrete wavelet transform (SID), g) morphological difference pyramid (MOD), and h) filter-subtract-decimate pyramid (FSD). Please refer to Figure 3 for an illustration of the fusion process, and also refer to [17] for an excellent description of these image fusion methods. In Section 5 we demonstrate that there is no universal fusion technique that is better for all the facial expressions. It was also reported in [18] that due to the subjective characteristics of the fusion performance evaluation, it is difficult to recommend a method for a given expression. However, in future we wish to investigate why a given fusion method performs better for a given expression.

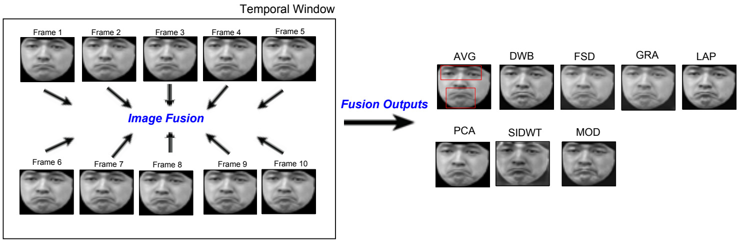


Fig. 3. Fusion of multiple images in a temporal window using i) AVG: local-average (the red bounding boxes illustrate regions where the local averaged were calculated from), ii) DWT: discrete wavelet transform , iii) FSD: filter-subtract-decimate pyramid, iv) GRA: gradient pyramid, v) LAP: laplacian pyramid, vi) PCA: principal component analysis, vii) SIDWT: shift invariant discrete (SID) wavelet transform, and viii) MOD: morphological difference pyramid

4 Experimental Setup

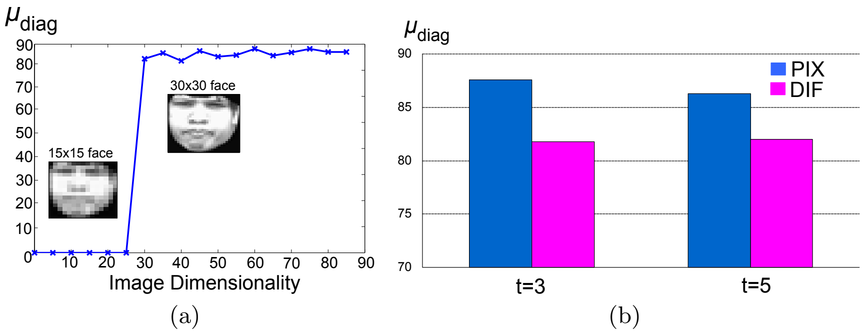
All experiments in this paper had been conducted with the objective of detecting the seven prototypic emotional facial expressions – anger, contempt, disgust, fear, happiness, sadness and surprise – which are available in the CK+ database. Active appearance models (AAMs) were employed for face-tracking, and its corresponding output SAPP pixel representations were used for training and testing the classifiers. For a fair comparison, the exact same two-fold cross validation train/test data partitions were adopted in all evaluations. A subject-independent approach was adopted in all evalutions. We shall adopt μ_{diag} to represent the weighted mean of the diagonal of the confusion matrix as the performance metric in all experiments.

4.1 AAM-Derived Pixel Representations

Active Appearance Models (AAMs) [19] have been shown to be a good method of aligning a pre-defined linear shape model that also has linear appearance variation, to a previously unseen source image containing the object of interest. In general, AAMs fit their shape and appearance components through a gradient-descent search. The shape, $\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} p_i \mathbf{s}_i$, of an AAM is described by a 2D triangulated mesh, which corresponds to a source appearance image; where $\mathbf{p} = (p_1, \ldots, p_m)^T$ are the shape parameters. In all our experiments, we report empirical results obtained from processing AAM-derived similarity normalized appearance features (i.e., SAPP pixel representations).

4.2 The Extended Cohn-Kanade Database

In this paper we used the Extended Cohn-Kanade (CK+) database [10], which contains 593 sequences from 123 subjects. The image sequences vary in duration (from 10 to 60 frames) and incorporate the onset (which is also the neutral frame) to peak formation of the facial expressions. For the 593 posed sequences, full FACS [20] coding of the peak frames had been provided.



(a) (b)

Fig. 4. (a) Once a threshold was exceeded (25 × 25 pixels), recognition rates were no longer significantly influenced by the image dimensionality. But, if image dimensionality fell below the threshold, then a unique solution to the objective function could not be found. (b) A deterioration in recognition rates was incurred when sparse absolute difference representations (DIF) were utilized in place of raw intensity pixel values (PIX).

5 Experimental Results

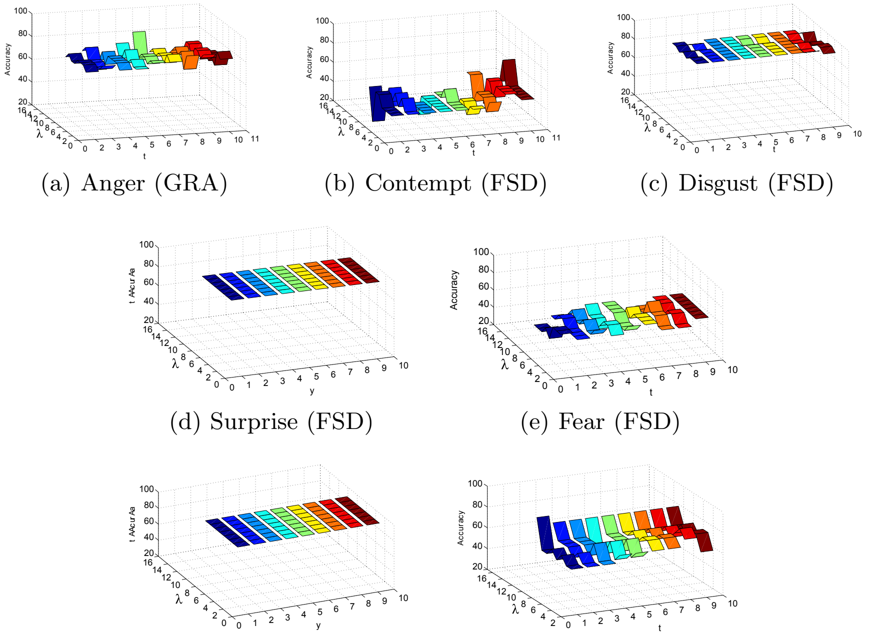
As mentioned in Section 3, we wish to first highlight the effect of naively concatenating temporal frames in ξ, such that $\widetilde{\xi} \in \mathbb{R}^{n\times(N\mapsto\tau)}$, is equivalent to the addition of noise; and therefore would have a detrimental effect on recognition rates. We further demonstrate that no benefits are introduced from using difference images due to a *lossy* effect inherent in subtractive operations on the holistic face. In fact, taking absolute difference images had produced substantial deterioration, which was mainly due to holistic information lost from the subtraction. These were supported by empirical results presented in Figure 4(b). In the SR classifier, the dimension of the random projection matrix Φ was taken to be a quarter that of the input image (we shall denote this by $\lambda = \frac{n}{d} = 4$).

5.1 Investigating Temporal Fusion and The Dimension of the Random Projection Matrix

In order to rectify the problems discussed in the previous section, two objectives must be fulfilled: i) the dimension N in ξ should not increase to obtain the sparsest $\vec{\alpha}^*$, and ii) holistic face information must be retained. These two criteria may be easily fulfilled through the utilization of image fusion techniques. All image fusion methods listed in Section 3 had been explored in our experiments. Concerning image dimensionality, we found that downsampling of facial images to a very low dimensionality was not suitable for expression recognition (unlike in face recognition [5]). Figure 4(a) shows that when all other variables except image dimensionality were held constant ($\lambda = 4$ and $t = 1$ in this experiment), the mean detection accuracy was not influenced once a threshold (25×25 pixels) on the image dimensionality was exceeded. In all subsequent experiments, the original image dimensionality had been preserved (87 × 93 pixels). It had also been observed that the dimension of the random projection matrix played a significant role in emotional expression detection. Denoting $\lambda = \frac{n}{d}$ as the factor by which the dimension of the projection matrix is downsampled with respect to the dimension of the input image, we were interested in analyzing the effect that varying the temporal window length t and λ had on recognition rates. The 3D plots (Accuracy versus time(t) versus λ) shown in Figure 5 shows that the optimal λ and t can be very different for different expressions. We observed that a larger t was more suitable for contempt, a larger λ was more suitable for sadness, and a larger λ coupled with a mid-range t was more suitable for anger. However, these two variables did not significantly influence happiness, disgust and surprise; which achieved near-perfect detections in our experiments.

5.2 Discussion

Recognition rates of the proposed temporal-SR classifier versus the static-SR classifier is presented in Table 1. For completeness, we have also included the performance of linear SVMs (which was trained and tested on in the exact same



(a) Anger (GRA)

(b) Contempt (FSD)

(c) Disgust (FSD)

(d) Surprise (FSD)

(e) Fear (FSD)

(f) Happiness (FSD)

(g) Sadness (AVG)

Fig. 5. 3D plots (X-axis (time t), Y-axis (random projection matrix dimension downsampling factor $\lambda = \frac{n}{d}$), Z-axis (Accuracy)) of the detection perfoamnces of the seven emotions as functions of time and λ. The image fusion method which produced the best recognition accuracy is shown in brackets.

Table 1. Recognition rates for emotion classification on the CK+ dataset for static-SR versus temporal-SR classification, and referenced to a linear SVM. μ_{diag} represents the weighted mean of the diagonal of the respective confusion matrices (computed through majority voting), and **N** represents the number of examples available from each emotion.

	N	static-SRC	temporal-SRC	SVM
Anger	45	90.9	95.5	86.1
Contempt	18	55.6	75.0	55.6
Disgust	59	100.0	100.0	100.0
Fear	25	66.7	75.0	91.7
Happiness	69	100.0	100.0	100.0
Sadness	28	85.7	92.9	85.7
Surprise	83	97.6	97.6	97.6
μ_{diag}	–	**91.9**	**94.9**	**93.2**

manner). As can be seen, the static-SR method experienced a deterioration of 1.3% with respect to the SVM, but once temporal information had been incorporated into the SR classifier, then a 3% improvement of the temporal-SR method over the static-SR method was afforded. Although it may appear on the surface that the differences between all three methods are not very significant, but we should not ignore the fact that the asympotote of ideal detection (i.e., perfect 100% recognition) is being approached and slight differences of a few percent may be more significant than as it would appear.

In view of this, it would be profoundly more interesting for an investigation to be conducted on more realistic facial expressions (i.e., acted and spontaneous) which possess deeper temporal dependencies for further insights of the underlying mechanisms of both static- and temporal-SR classifiers to be gained. In addition, since SVMs have been actively employed in expression recognition, it would also be interesting to make a direct comparison with the SR classifiers by employing fused temporal information. Such an analysis would stimulate an interesting thought-provoking analysis on which is more capable at exploiting expression dynamics – the ℓ_1-norm or the ℓ_2-norm?

6 Conclusion and Future Work

In this paper, we explored the method of sparse representation (SR) classification to detect the seven prototypic emotion-related facial expressions. Having established the importance of expression dynamics, we proposed a framework in which a dynamic model could be effectively implemented into the SR classifier. Our work explored the logic behind the use of sparse features in the SR framework, and also investigated the influence of the dimensions of the random projection matrix and length of the temporal window. Indeed, we found that the latter two were significant factors in influencing detection performance. Additionally, various techniques of incorporating temporal information into feature matrix ξ had been analyzed and proposed. In future work, we intend to investigate if the dynamics of more realistic and spontaneous facial expressions (on both emotional-related expressions and action units) could be exploited using our proposed method. Apart from this, we wish to analyze in further detail why the filter-subtract-decimate pyramid image fusion method was more suitable for most expressions, but not for the remaining few.

Acknowledgments. This research was supported in part by the Cooperative Research Centre for Advanced Automotive Technology (AutoCRC).

References

1. Kobayashi, H., Hara, F., Ikeda, S., Yamada, H.: A basic study of dynamic recognition of human facial expressions. In: 2nd IEEE International Workshop on Robot and Human Communication, pp. 271–275 (1993)
2. Pantic, M., Patras, I.: Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 4, pp. 3358–3363 (2005)
3. Valstar, M., Pantic, M.: Fully automatic facial action unit detection and temporal analysis. In: Computer Vision and Pattern Recognition Workshop CVPRW 2006, pp. 149–149 (2006)
4. Tong, Y., Liao, W., Ji, Q.: Facial action unit recognition by exploiting their dynamic and semantic relationships. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 1683–1699 (2007)
5. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2008)
6. Bociu, I., Pitas, I.: A new sparse image representation algorithm applied to facial expression recognition. In: Proceedings of the 14th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, 2004, p. 539 (2004)
7. Zafeiriou, S., Petrou, M.: Sparse representations for facial expressions recognition via l1 optimization. In: Computer Vision and Pattern Recognition Workshops (CVPRW), p. 32 (2010)
8. Cotter, S.: Recognition of occluded facial expressions using a fusion of localized sparse representation classifiers. In: Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE), p. 437 (2011)
9. Mahoor, M., Zhou, M., Veon, K.L., Mavadati, S., Cohn, J.: Facial action unit recognition with sparse representation. In: Automatic Face & Gesture Recognition and Workshops, p. 336 (2011)
10. Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: Proceedings of the IEEE Workshop on CVPR for Human Communicative Behavior Analysis (2010)
11. Yang, P., Liu, Q., Cui, X., Metaxas, D.: Facial expression recognition using encoded dynamic features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
12. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 915–928 (2007)

13. Du, R., Wu, Q., He, X., Jia, W., Wei, D.: Facial expression recognition using histogram variances faces. In: 2009 Workshop on Applications of Computer Vision, WACV (2009)
14. Ying, Z.-L., Wang, Z.-W., Huang, M.-W.: Facial Expression Recognition Based on Fusion of Sparse Representation. In: Huang, D.-S., Zhang, X., Reyes García, C.A., Zhang, L. (eds.) ICIC 2010. LNCS, vol. 6216, pp. 457–464. Springer, Heidelberg (2010)
15. Candes, E.: The restricted isometry property and its implications for compressed sensing. Comptes Rendus Mathematique 346, 589–592 (2008)
16. Hubel, D.: Eye, Brain and Vision. Freeman (1987)
17. Rockinger, O., Fechner, T.: Pixel-level image fusion: The case of image sequences. In: Proc. SPIE, vol. 3374, pp. 378–388 (1998)
18. Canga, E.F.: Image fusion. Project report, Dept. Electronic and Electrical Eng., Univ. of Bath. (2002)
19. Cootes, T., Edwards, G., Taylor, C.: Active Appearance Models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 681–685 (2001)
20. Ekman, P.: Emotion in the human face. Cambridge University Press (1982)