Improved Facial Expression Recognition via Uni-Hyperplane Classification

S.W. Chew*, S. Lucey†, P. Lucey‡, S. Sridharan*, and J.F. Cohn‡

Abstract

Large margin learning approaches, such as support vector machines (SVM), have been successfully applied to numerous classification tasks, especially for automatic facial expression recognition. The risk of such approaches however, is their sensitivity to large margin losses due to the influence from noisy training examples and outliers which is a common problem in the area of affective computing (i.e., manual coding at the frame level is tedious so coarse labels are normally assigned). In this paper, we leverage the relaxation of the parallel-hyperplanes constraint and propose the use of modified correlation filters (MCF). The MCF is similar in spirit to SVMs and correlation filters, but with the key difference of optimizing only a single hyperplane. We demonstrate the superiority of MCF over current techniques on a battery of experiments.

1. Introduction

Research into affective computing has been very active over the past decade, mainly driven by social, economic and commercial interests (such as behavioral science, human-computer-interaction, health-care, security, etc). The main goal of this research is to have a computer system being able to automatically detect/infer the emotional state of any person based on various modes (e.g., face, voice, body, actions) in real-time.

The plurality of this research has been anchored in facial expression recognition, which consists generally of tracking/registering faces automatically and then training corresponding models for supervised classification. To accomplish the former, one may opt for a coarse form of face registration (e.g., Viola-Jones [4]) followed by a mapping to higher-dimensional feature spaces (e.g., Gabor magnitudes [5], local binary pattern operators [6], histogram of oriented gradients [7], etc); or choose an algorithm (e.g., active appearance models [8], constrained local models [9], etc) which administers more precise registration but allows the feature extraction step to be bypassed (if illumination conditions are known to be consistent). Whichever the choice, both avenues ultimately lead to the final pattern classification stage, which is usually realized in the literature through support vector machines (SVM).

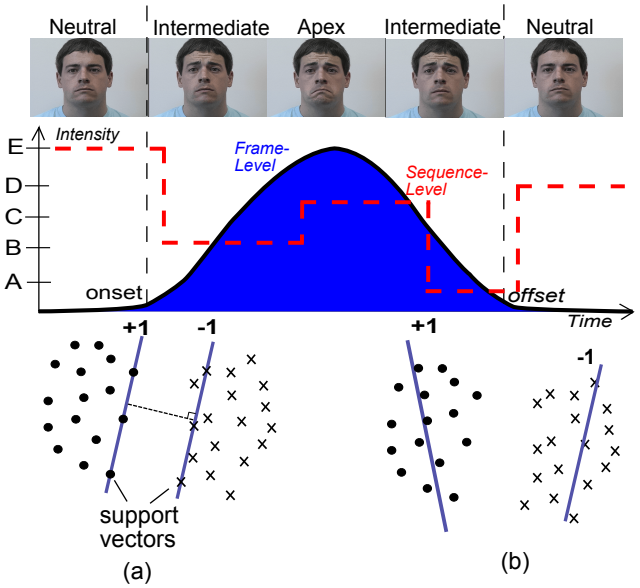

Figure 1. Given labels at the frame-level, facial expression recognition may be accomplished relatively accurately, but this accuracy rapidly declines once coarse labels which produce ambiguous codes are introduced at the sequence-level. (a) Support vector machines require labels which are precisely defined, and therefore do not adapt well to much noisier coarse labels because they require both hyperplanes to be (i) strictly parallel to each other, and (ii) at a tangent to the support vectors. (b) By relaxing one or both these requisites (e.g., in PSVM [1], GEPSVM [2] and TWSVM [3]), the resulting classifier can be accommodated to be much more robust against noisy and outlying training examples.

Unfortunately, the SVM still bears several imperfections in its traditional formulation. Numerous modifications had been proposed over the years, but in spite of all these encouraging recommendations, there remains a

*S.W. Chew and S. Sridharan are with the SAIVT lab at Queensland University of Technology, Brisbane, Australia. Email: {sien.chew@qut.edu.au, s.sridharan@qut.edu.au}.

†S. Lucey is with the Commonwealth Science and Industrial Research Organisation (CSIRO), Australia; Email: {simon.lucey@csiro.au}.

‡P. Lucey is with Disney Research Pittsburgh and J.F. Cohn is with the Department of Psychology, University of Pittsburgh/Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 15260. Email: {patrick.lucey@disneyresearch.com, jeffcohn@cs.cmu.edu}.

lingering concern regarding the fundamental geometrical reasoning which demands that both hyperplanes are to be strictly parallel and at a tangent to the support vectors. Fung and Mangasarian [1] proposed proximal support vector machines (PSVM) which abolishes the latter rule, and as a consequence managed to deliver impressive improvements. Generalized eigenvalue proximal support vector machines (GEPSVM) [2] and twin support vector machines (TWSVM) [3], are two notable advancements to the PSVM which abolish both rules altogether. These valuable contributions all persuades one to break free from the traditional "parallel-hyperplanes" train-of-thought (see Fig. 1), and begs the question – why is the requirement for two hyperplanes always necessary?

In this paper, our goal is to improve generalization performance by optimizing only a single hyperplane, that of which corresponds to the positive set. Intuitively, the advantages of doing so are twofold – a) the trade-off resulting from the joint-optimization of two hyperplanes is diminished, and b) outliers in the negative set are no longer allowed to significantly influence the positioning of the positive hyperplane. The proposed algorithm, which we refer to as "modified correlation filters" (MCF), may be considered very similar in principle to correlation filters, but closely-related mathematically to SVMs.

The central contributions from this paper are,

- Propose a novel supervised classifier (MCF) which is inspired by correlation filters and similar in spirit to support vector machines, but the key difference being that it functions with only a single hyperplane.
- Show that MCF exhibits superior robustness against noisy/outlying training examples, and thus enable significant improvements in generalization performance to be gained.
- Demonstrate the benefits of MCF on a variety of challenging facial expression databases, which include posed, acted and spontaneous expressions.

2. Related Work

2.1. Classifiers Utilized in Expression Recognition

At the discipline's infancy, a significant amount of effort had been expended into establishing which classification algorithm was most suitable for expression recognition. Classifiers that had generated considerable excitement include [10] artificial neural networks, linear discriminant analysis, hidden Markov models, dynamic Bayesian networks and various expert rules. Following the seminal work of [11], however, a consensus was generally agreed upon that soft-margin support vector machines (SVM hereon) were deemed to be more applicable for expression recognition problems, mainly due to the following properties [11]: i) good generalization performance, ii) capitalizing well on well-correlated input spaces (a characteristic of facial expressions), and iii) relatively short training times $\mathrm{O}(10^2)$ required.

Enthusiasm into classifier research had since steadily declined. Recently, the greater part of research interests had been proposed in favour of the classifier, suggesting that a large number of researchers appear generally satisfied with SVMs. Much of the attention that was diverted away from the classifier appear converged onto interests such as: i) minimizing the registration error resulting from face-tracking [12, 13], ii) exploring different forms of feature representations [7, 5, 14], and iii) modelling space-time relationships of various facial components [14, 15]. Even though these research concerns are amply justified, but the relevance of pursuing new suitable classifiers should not be undermined because final decisions are ultimately determined at the classification stage. Before proceeding to SVM theory, we shall discuss shortly on correlation filters which was instrumental in motivating our proposed MCF.

2.2. Correlation Filters

Fundamental principles which had aided in the development of MCF were originally drawn from correlation filters, and as such, due acknowledgments shall be accorded here. However, it should be mentioned that the similarity between these algorithms exist solely in terms of philosophy, and not mathematically. Using a variety of cross-correlation methods, correlation filters are generally employed to detect the presence of a pattern in a target (e.g., face verification [16], visual speaker verification [17], etc). One usually applies the Fourier Transform to speed up computation. For example, the *Minimum Average Correlation Energy (MACE)* filter [18] attempts to emphasize the correlation peak by minimizing the average correlation plane energy, and hence enhancing the peak-to-sidelobe ratio.

A shared objective between correlation filters and MCF lies in minimizing the energy of non-targets (i.e., negative examples) in order to emphasize the targets (i.e., positive examples). To the best of the authors' knowledge, there had been no prior application of correlation filters to facial expression recognition problems, probably due to their poor generalization capability. Nonetheless, the empirical performance of MACE filters ($\mathbf{H} = D^{-1}X(X^{+}D^{-1}X)^{-1}\mathbf{u}$) [18] shall be presented alongside those of the SVM and MCF in §6. As shall be discussed in §4, the mathematics behind MCF and correlation filters are shown to be entirely unrelated.

3. Evolution of the SVM

In contrast to correlation filters, SVMs have played an instrumental role in the facial expression recognition literature. In this section, we review SVMs from a probabilistic

perspective, and then use the ensuing insights to motivate our proposed MCF in §4. Traditional approaches to supervised classification consider the following: in a binary problem where we are given two sets of patterns, the goal is to assign an unknown test pattern to one of the two sets,

$$(\mathbf{x}_i, \mathbf{y}_i) \in \chi \times \{\pm 1\}. \tag{1}$$

Here, the objective is to learn the mapping $f : \chi \mapsto \{\pm 1\}$, where we assume that X and Y are two sets of random variables, and χ consists of m samples which are drawn iid (independently and identically distributed) from the probability distribution $X \times Y$. This suggests that the random variables must be generated from a *fixed* but *unknown* probability distribution $p(x, y)$ from either one set or the other,

$$p(x_i, y_i|f) = \prod_{i=1}^{m} p(y_i, x_i|f)p(x_i). \tag{2}$$

The problem of achieving good generalization can then be cast as learning a mapping f_χ which reduces the probability of obtaining incorrect label assignments by as much as possible [19],

$$z_i := Pr(sign(f_\chi(\mathbf{x}_i))) \neq y_i. \tag{3}$$

(3) can be equivalently interpreted as minimizing the expected or empirical risk,

$$R_{exp}[f] = \int_{X \times Y} \lfloor \mathbf{z} + 0.5 \rfloor \quad dP, \tag{4}$$

$$R_{emp}[f] = \frac{1}{m} \sum_{i=1}^{m} \lfloor z_i + 0.5 \rfloor. \tag{5}$$

In practice, $R_{emp}[f]$ is usually employed as $R_{exp}[f]$ is an intuitive quantity; but (5) still manifests as an ill-posed problem due to the set of functions $\{f_\chi\}$ which remain unknown. To regularize this, an upper-bound on $R_{emp}[f]$ is enforced on $R_{emp}[f]$, and a subset of feasible regions in f_χ is constrained to lie within $||f||_K$ and bounded by R (this method is commonly referred to as Ivanov regularization),

$$\min_{f_\chi \in H, ||f_\chi||_K \leq R} \quad \frac{1}{m} \sum_{i=1}^{m} \lfloor (f(x_i) \neq y_i) + 0.5 \rfloor, \tag{6}$$

where K is a positive-definite kernel function [20] living in a Reproducing Kernel Hilbert Space. Even though the formulation now appears better defined, the optimization problem (6) is still NP-complete [19], and therefore intractable as a result of the non-smooth and non-convex summands (which are essentially zero-one losses). A remedy for this involves replacing all these loss terms by a smooth and convex loss function $V(y, f(\mathbf{x_i}))$ while still enforcing an upper bound on $R_{emp}[f]$. Again, it is not always tractable in practice for a smooth $V(y, f(\mathbf{x_i}))$ to co-exist in (6). Instead of minimizing $R_{emp}[f]$ subject to strict bounds on $||f||_K^2$, it is far more tractable to smoothly trade-off $||f||_K^2$ and $R_{emp}[f]$; thus giving rise to the closely-related Tikhonov regularization problem,

$$\min_{f \in H} \quad \frac{1}{m} \sum_{i=1}^{m} V(y_i, f(\mathbf{x}_i)) + \lambda ||f||_K^2. \tag{7}$$

Here, λ effectively controls the regularization trade-off between $||f||_K^2$ with $R_{emp}[f]$. SVMs are often interpreted as a form of Tikhonov regularization problem expressed in the form of (7) where,

$$V(y_i, f(\mathbf{x}_i)) = [1 - y_i \mathbf{w}^T h(\mathbf{x}_i)]_+, \tag{8}$$

$$||f||_K^2 = ||\mathbf{w}||_p^p, \tag{9}$$

and $[f(\mathbf{z})]_+ = \max(0, f(\mathbf{z}))$, (8) and (9) represent the hinge and margin loss terms respectively, and $p \in \{1, 2\}$. A typical choice of SVM in automatic facial expression recognition problems is the linear ℓ_2-SVM (i.e., $p = 2$), mainly because of their rapid training times and good generalization capabilities,

$$\arg\min_{\mathbf{w}, b} \quad \frac{1}{m} \sum_{i=1}^{m} [1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)]_+ + \frac{\lambda}{2} ||\mathbf{w}||_2^2. \tag{10}$$

3.1. A Limitation of SVMs

One of the inherent drawbacks of SVM lies in its sensitivity to large margin losses due to the influence from noisy examples and outliers. Firstly, [21] pointed out that it is not a trivial task for overlapping distributions to be modelled. This is mainly due to a lack of probabilistic insight into the margin loss $\lambda ||f||_H$, where an intuitive procedure of tuning λ still does not yet exist. However, [21] and [22] did point out that this inadequacy may be redressed to some degree by enforcing an upper bound on the slack variables, so as to diminish the influence from extreme outliers in the training set (since these outliers contribute significantly to the largest margin loss).

Similarly, other researchers had looked at working with the non-twice-differentiable (and hence non-convex) property of the hinge-loss functional. In [23], non-convexity was dealt with through a multi-stage relaxation of the hinge-loss using semi-definite programming methods. Others had proposed to replace the hinge-loss altogether with – a smooth sigmoid function [24], a least-median-loss function [25] and a non-linear Gaussian error function [26] to depreciate the leverage that extreme outliers hold on the margin loss.

From this plethora of literature, we observed one commonality shared among these modifications was that most had been directed at the loss-functional $V(y_i, f(\mathbf{x}_i))$. Equations (1) to (9) had illustrated how SVMs may be interpreted

as a Tikhonov regularization problem that attempts to minimize the empirical risk (which is equivalent to the loss-functional). From these perspectives, we appreciate the importance of the loss-functional's role in the objective function, and its significance in filtering out outliers in χ. This intuition forms a primary motivation for MCF.

3.2. Recent Advancements of SVMs

More recently, these machines have subtly but clearly evolved into different embodiments of their original form. Keen interest was noted in disputing the elementary axioms of the SVM, which require: i) hyperplanes to be parallel, and ii) hyperplanes to be positioned at the support vectors. Proximal-SVMs (PSVM) [1] defy the latter rule,

$$\begin{aligned} \arg\min_{\mathbf{w},b} \quad & \frac{\nu}{2}||\delta_i||_2^2 \quad + \quad \frac{1}{2}(\mathbf{w}^T\mathbf{w} + b^2) \\ & s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1 - \delta_i. \end{aligned} \tag{11}$$

where b refers to the bias, and δ_i refers to the slack variables. It is worth noting that this was achieved by transforming $V(y_i, f(\mathbf{x}_i))$ from an inequality constraint into an equality constraint. Here, the equality sign allows the parallel bounding planes to "break free" from the support vectors and to be pushed as far apart from each other.

Generalized eigenvalue proximal SVMs (GEPSVM) [2] and Twin-SVMs (TWSVM) [3] are algorithms that have recently evolved from the PSVM. Both inherited the PSVM's concept of 'hyperplane repulsion', but the parallelism rule had been further dropped.

4. Modified Correlation Filter

PSVM, GEPSVM and TWSVM all suggest an alternate perspective in how SVMs may be formulated. These modifications, as well as those described in §3.1, all appear to place important emphasis on $V(y_i, f(\mathbf{x}_i))$. A key concept behind the modified correlation filter (MCF) had emanated from similar standpoints, and then merged with the philosophy behind correlation filters (see Fig. 2).

Instead of relying on error-loss variables from both positive and negative sets to define $V(y_i, f(\mathbf{x}_i))$, MCF uses only error variables from the positive set,

$$\mathbf{w}^* = \arg\min_{\mathbf{w},b} \quad \frac{1}{m}\sum_{i=1}^{m}[1 - \mathbf{w}^T\mathbf{x}_i^{(+)} + b]_+ + \frac{\lambda}{2}\mathbf{w}^T\mathbf{Q}\mathbf{w}, \tag{12}$$

$$\mathbf{Q} = [\mathbf{X} \quad \mathbf{e}][\mathbf{X} \quad \mathbf{e}]^T, \tag{13}$$

$$\mathbf{X} = [\mathbf{X}^{(+)} \quad \mathbf{X}^{(-)}]. \tag{14}$$

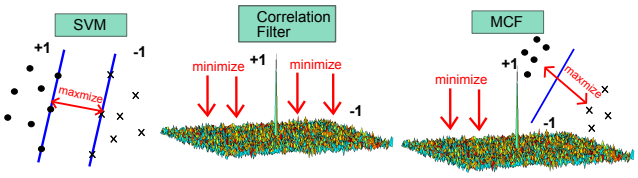

Figure 2. An intuitive comparison between SVM, correlation filters and MCF. (a) SVM computes the widest margin between classes in order to detect '+1', (b) Correlation filters minimizes the correlation plane energy of '-1' in order to emphasize the correlation '+1' peak. (c) MCF had drawn inspiration from both SVM and correlation filters, and operates by seeking to emphasize '+1' by minimizing the energy of *all* examples except '+1' examples, as cast in a SVM optimization framework.

Here, the goal is learn an optimal uni-hyperplane classifier $\mathbf{w}^* \in \mathbb{R}^n$, from positive and negative sets which are represented by $\mathbf{X}^{(+)} \in \mathbb{R}^{n\times|\mathbf{X}^{(+)}|}$ and $\mathbf{X}^{(-)} \in \mathbb{R}^{n\times|\mathbf{X}^{(-)}|}$ respectively, b refers to the bias, $\lambda > 0$, $m = |\mathbf{X}^{(+)}|$, $\mathbf{e} \in \mathbb{R}^n$ indicates a vector of ones, and $|\cdot|$ refers to set cardinality. Note that in (12), *only* examples from the positive set reside in $V(y_i, f(\mathbf{x}_i))$. There are several advantages in doing so. First and foremost, the number of outliers in χ required for optimization are considerably decreased (since $|\mathbf{X}^{(+)}| << |\mathbf{X}^{(+)}|$), and therefore the empirical risk is in turn reduced. In order to distinguish positive from negative examples, we insert both positive and negative sets into the precomputable matrix $\mathbf{Q}$ in (13). From a Tikhonov regularization standpoint, this may be interpreted as a trade-off between empirical risk and a margin loss that is defined upon the energy of all training examples in χ. It is in this respect that MCF share a common philosophy with correlation filters, but note that the similarity does not extend to their mathematical definitions.

The Lagrangian corresponding to the problem (12) is given by,

$$\begin{aligned} L(\mathbf{w}, b, \delta, \alpha, \beta) = & \frac{1}{2}[\mathbf{X}\mathbf{w} + \mathbf{e}_1 b]^T[\mathbf{X}\mathbf{w} + \mathbf{e}_1 b] + \lambda\mathbf{e}_2\delta \\ & -\alpha^T[\mathbf{X}^{(+)}\mathbf{w} + \mathbf{e}_2 b + \delta - \mathbf{e}_2] - \beta^T\delta, \end{aligned} \tag{15}$$

where α and β represent the Lagrange multipliers, and $\mathbf{e}_1$ and $\mathbf{e}_2$ are vectors of ones of appropriate dimensions. The Karush-Kuhn-Tucker [27] necessary and sufficient optimality conditions are given by,

$$\mathbf{X}^T[\mathbf{X}\mathbf{w} + \mathbf{e}_1 b] - \mathbf{X}^{(+)T}\alpha = 0, \tag{16}$$

$$\mathbf{e}_1^T[\mathbf{X}\mathbf{w} + \mathbf{e}_1 b] - \mathbf{e}_2^T\alpha = 0, \tag{17}$$

$$\lambda\mathbf{e}_2 - \alpha - \beta = 0, \tag{18}$$

$$\mathbf{X}^{(+)}\mathbf{w} + \mathbf{e}_2 b + \delta \geq \mathbf{e}_2, \tag{19}$$

Figure 3. Experiments were conducted on a wide variety of facial expression databases: (a) CK+ dataset (posed expressions), (b) UNBC-McMaster Shoulder Pain Archive (spontaneous expressions), and (c) GEMEP-FERA dataset (acted expressions).

$$\alpha^T(\mathbf{X}^{(+)}\mathbf{w} + \mathbf{e}_2 b) + \delta - \mathbf{e}_2 = 0; \quad \beta^T\delta = 0, \tag{20}$$

$$\alpha \geq 0, \quad \beta \geq 0. \tag{21}$$

From (18) and (21) we have,

$$0 \leq \alpha \leq \lambda. \tag{22}$$

By combining and then simplifying (16) and (17),

$$[\mathbf{X}^T \quad \mathbf{e}_1^T][\mathbf{X} \quad \mathbf{e}_1][\mathbf{w}, b]^T - [\mathbf{X}^{(+)T} \quad \mathbf{e}_2^T]\alpha = 0, \tag{23}$$

$$\mathbf{G} := [\mathbf{X} \quad \mathbf{e}_1], \ \mathbf{H} := [\mathbf{X}^{(+)} \ \mathbf{e}_2], \ \mathbf{u} := [\mathbf{w}, b]^T, \tag{24}$$

$$\mathbf{u} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{H}^T\alpha. \tag{25}$$

Using (25) and the KKT conditions (16-21) we obtain the Wolfe Dual of MCF as,

$$\arg\max_{0\leq\alpha\leq\lambda} \quad \mathbf{e}_1^T\alpha - \frac{1}{2}\alpha^T\mathbf{H}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{H}^T\alpha. \tag{26}$$

4.1. Robust Primal Form

By relinquishing a need for the joint-optimization of two sets, there is little justification for choosing the value '1' as reference points in the hinge-loss functional (unlike in the SVM). Instead, we can replace the constant with an offset parameter ρ,

$$\mathbf{w}^* = \arg\min_{\mathbf{w},b} \quad \tfrac{1}{m}\textstyle\sum_{i=1}^{m}[[\rho - \mathbf{w}^T\mathbf{x}_i^{(+)} + b]_+]_- + \tfrac{\lambda}{2}\mathbf{w}^T\mathbf{Q}\mathbf{w}, \tag{27}$$

$$\rho = 1 - \text{sign}(\min\{\mathbf{w_0}^T\mathbf{x}_i^{(+)}\}_{i=1}^m). \tag{28}$$

where $[f(\mathbf{z})]_- = \min(\rho, f(\mathbf{z}))$. Because ρ ensures that $[f(\mathbf{z})]_+$ is never allowed to be negative during the first iterate, all examples in $\mathbf{X}^{(+)}$ are guaranteed to be optimized. Furthermore, $[f(\mathbf{z})]_-$ diminishes the influence from extreme outliers by upper-bounding the maximum value of the slack variables to be at most ρ.

5. Facial Expression Databases

All experiments conducted in this paper were designed to evaluate all classifiers under study for the tasks of frame-level action unit (AU) detection and sequence-level emotion-related expression detection. These experiments were designed to: (i) evaluate the classification response of MCF, and (ii) reference the recognition accuracy of MCF with respect to SVM[1] and correlation filters[2]. In this section, we describe three facial expression databases (Fig. 3) which had been carefully selected to reflect varying levels of noise conditions usually encountered in practice.

5.1. The Extended Cohn-Kanade database (CK+)

The CK+ database [28] consists of 593 FACS coded sequences from 123 subjects eliciting posed facial expressions. The image sequences vary in duration (from 10 to 60 frames) and incorporate the onset to peak formation of the facial expressions. In our experiments, we focused on the following AUs: {1 2 4 6 7 12 15 17 25 26}, and all seven emotions (i.e., anger, contempt, disgust, fear, happiness, sadness and surprise).

5.2. The GEMEP-FERA Database

The GEMEP-FERA database [6] contains recordings of 10 actors expressing a total of 15 emotions together with a variety of AUs which had been FACS coded. In all of these recordings, actors were instructed to utter meaningless phrases (such as the sustained vowel 'aaa') with the

[1] We used linear SVM, PSVM and GEPSVM in all of our experiments.
[2] We used MACE filters in all of our experiments.

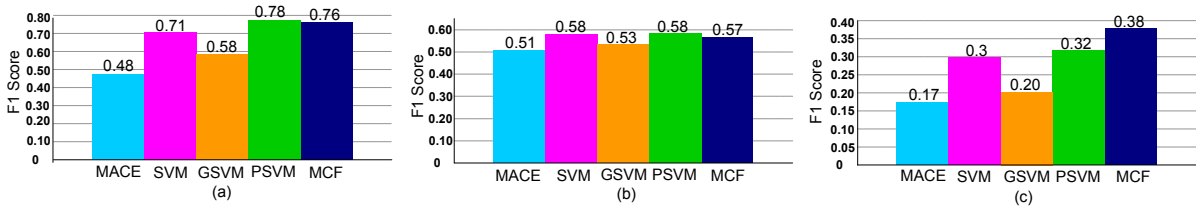

Figure 4. AU detection performance of MACE, SVM, GEPSVM, PSVM and MCF on: (a) CK+ dataset, (b) GEMEP-FERA dataset, and (c) Pain-AU dataset. Performance was evaluated using the weighted F1-score, which are indicated numerically as shown.

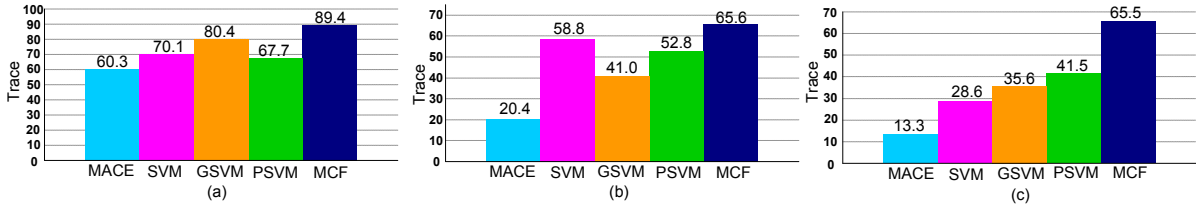

Figure 5. Emotion detection performance of MACE, SVM, GEPSVM, PSVM and MCF on: (a) CK+ dataset, (b) GEMEP-FERA dataset, and (c) Pain-OPR dataset. Performance was evaluated using confusion matrices computed through majority voting. The trace of the respective confusion matrices are indicated numerically as shown (details for the Pain-OPR intensity confusion matrices can be found in Fig. 6).

aid of a professional director. The key difference between this dataset with the CK+ and UNBC-McMaster Shoulder Pain Archive is that expressions had been displayed in the presence of speech, which generated a substantial amount of rigid head and body motion. In our experiments, we focused on the following AUs: {1 2 4 6 7 10 12 15 17 18 25 26}, and all five emotions (i.e., anger, fear, joy, relief and sadness).

5.3. The UNBC-McMaster Shoulder Pain Archive

The UNBC-McMaster Shoulder Pain Expression Archive [29] (Pain dataset hereon) contains 200 video sequences spanning 25 subjects which were recorded of their faces while they moved their affected (these subjects had various shoulder injuries) and unaffected shoulders. Characteristic of spontaneous facial expressions, the video sequences have various durations (from 90 to 700 frames) in which considerable head movement had been exhibited. In the AU portion of our experiments, we focused on the following AUs: {4 6 7 9 10 12 25 26 43}. Heavy emphasis had been placed on the task of distinguishing between observer-rated pain-intensity levels, which were based on a six-point scale – from OPR0 to OPR5, in increasing levels of observed pain intensity.

6. Experiments

As generalization performance was our principal matter of interest in this paper, a leave-one-subject-out cross-validation procedure had been adopted in all six experiments[3]: i) AU and emotion detection on the CK+ dataset, ii) AU and emotion detection on the GEMEP-FERA dataset, iii) AU and pain-intensity detection on the Pain dataset hereon. Normalized pixel representations, acquired using subject-independent constrained local models (CLM) [9], were employed as input into the classifiers. Specifically, the pixel representations employed in this study are referred to as canonical normalized appearance features (CAPP). The only exception was in the Pain-OPR experiments where substantial difficulty in achieving even reasonable levels of detection was experienced. For this experiment only, we employed AAM-derived CAPP representations instead. It was postulated in [12, 30] that no significant benefits could be obtained from utilizing appearance features once close to ideal registration had been obtained under consistent illumination conditions (which is valid for all datasets under analysis here). Hence, we had carefully neglected the use of appearance features in our experiments.

6.1. Frame-level AU Detection

Of all classifiers examined, MACE had exhibited the poorest performance in all AU experiments (Fig. 4). Considering that correlation filters are inherently subject-dependent by design, and because the filters had been evaluated subject-independently, this result was not unusual.
MCF offered significant improvements over SVM on the CK+ and Pain datasets, but equal levels in performance on the GEMEP-FERA dataset. GEMEP-FERA contains only seven subjects, therefore implying lower inter-subject variability (i.e., noise and outliers) compared to the other two datasets. Interestingly, MCF still had not performed worse than SVM. PSVM proved well-suited for AU detection as noted by its good performance on all three AU datasets. On

[3] All classifiers had been evaluated in the exact same manner.

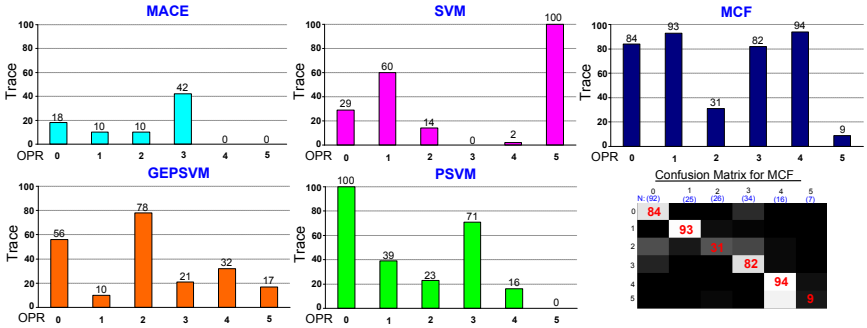

Figure 6. The diagonal elements of the respective confusion matrices obtained for Pain-OPR intensity-level detection experiments are illustrated . N indicates the number of examples available for each sequence. The traces of these respective confusion matrices can be found in Fig. 5(c).

the other hand, GEPSVM had not performed too well. One possible explanation for this could be the sensitivity of its non-parallel hyperplanes towards position and orientation, which may have been adversely affected by the presence of several different AUs that were included in the negative training set.

6.2. Sequence-based Emotion Detection

The best classification rates were achieved by MCF (Fig. 5). As with AU detection, similar trends in performance were observed in MACE, SVM and PSVM, except for GEPSVM on the CK+ dataset. Here, the construction of "non-parallel-hyperplanes" appeared more applicable for the task of distinguishing between emotions, but only if the data was relatively noise-free (posed expressions). This observation, however, became less apparent in the "noisier" GEMEP-FERA dataset.

6.3. Intensity-based Experiments

The most interesting experiment tasked all classifiers to identify between six different levels of pain-intensity in the Pain dataset. The high difficulty of this task lies with the spontaneity of the expressions, marked by substantially increased levels of both intra-/ and inter-subject expression variability. Most noticeably, a large component of the difficulty is attributed to the visual similarity between adjacent intensity levels (e.g., OPR 0 contained numerous frames which appeared very similar to OPR 1). Hence, there was considerable overlap between positive and negative training examples; and therefore provoking large numbers of outliers in χ.

As observed in Fig. 5(c), MCF provided the best performance, delivering a two-fold improvement over SVM. PSVM and GEPSVM both performed better than SVM. Even so, the difference between MCF and the next-best classifier (PSVM) was still observed to be very significant. Details of the respective confusion matrices are presented in Fig. 6. Referring to the last row of the MCF confusion matrix (corresponding to OPR 5), there was some difficulty encountered by MCF at distinguishing OPR 5 from OPR 4. It should be emphasized again that both these intensity-levels in close proximity to one another appear very similar even visually. Equally important, there were only very few (seven) examples for OPR 5 available for training. In spite of all these, OPR 5 had not been misclassified by MCF to be of lower-intensity levels (i.e., OPR 0 to 3).

7. Conclusion

We propose MCF, a supervised binary classification algorithm inspired by correlation filters and SVMs. MCF conserves the energy of the target using linear constraints, and then minimizes the energy of both targets and non-targets in the objective function. Doing so effectively reduces the influence of outliers and noise in the training set. We demonstrated the advantages of MCF on automatic facial expression recognition problems, but we point out that MCF may also be applied to other similar classification problems. Rigorous analysis was conducted through a battery of posed, acted and spontaneous facial expression recognition experiments, which had demonstrated the usefulness of MCF over SVM (plus several of its variants) and correlation filters.

8. Acknowledgments

This research was supported in part by the Cooperative Research Centre for Advanced Automotive Technology (AutoCRC) and the National Institute of Mental Health grant R01 MH51435. The authors gratefully acknowledge the contribution of National Research Organization and reviewers' comments.

References

[1] G. Fung and O. Mangasarian, "Proximal support vector machine classifiers," in *Proc. Knowledge Discovery and Data Mining*, F. Provost and R. Srikant, Eds., 2001, pp. 77–86. 1, 2, 4

[2] O. Mangasarian and E. Wild, "Multisurface proximal support vector machine classification via generalized eigenvalues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 69–74, Jan. 2006. 1, 2, 4

[3] Jayadeva, R. Khemchandani, and S. Chandra, "Twin support vector machines for pattern classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 905–910, May 2007. 1, 2, 4

[4] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518. 1

[5] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic Recognition of Facial Actions in Spontaneous Expressions," *Journal of Multimedia*, 2006. 1, 2

[6] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The First Facial Expression Recognition and Analysis Challenge," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. 1, 5

[7] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2005. 1, 2

[8] T. Cootes, G. Edwards, and C. Taylor, "Active Appearance Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001. 1

[9] J. Saragih, S. Lucey, and J. Cohn, "Face Alignment through Subspace Constrained Mean-Shifts," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009. 1, 6

[10] B. Fasel and J. Luettin, "Automatic Facial Expression Analysis: A Survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003. 2

[11] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," in *Computer Vision and Pattern Recognition Workshop, CVPRW*, Jun. 2004, p. 80. 2

[12] S. Chew, P. Lucey, S. Lucey, J. Saragih, J. Cohn, and S. Sridharan, "Person-independent facial expression detection using constrained local models," in *IEEE Workshop on Facial Expression Recognition and Analysis Challenge, at AFGR*, 2011. 2, 6

[13] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using phog and lpq features," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 878–883. 2

[14] B. Jiang, M. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 314–321. 2

[15] T. Wu, M. Bartlett, and J. Movellan, "Facial expression recognition using gabor motion energy filters," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 2010, pp. 42–47. 2

[16] M. Savvides, B. Vijaya Kumar, and P. Khosla, "Face verification using correlation filters," in *Proceedings of the Third IEEE Automatic Identification Advanced Technologies*, March 2002, pp. 56–61. 2

[17] D. Ramli, S. Samad, and A. Hussain, "A umace filter approach to lipreading in biometric authentication system," *Journal of Applied Sciences*, vol. 8, pp. 280–287, 2008. 2

[18] A. Mahalanobis, B. Kumar, and D. Casasent, "Minimum average correlation energy filters," *Applied Optics*, vol. 26, no. 17, pp. 3633–3640, Sep 1987. 2

[19] R. Rifkin, G. Yeo, and T. Poggio, "Regularized least-squares classification," *Nato Science Series Sub Series III Computer and Systems Sciences*, vol. 190, pp. 131–154, 2003. 3

[20] G. Wahba, "Support vector machines, reproducing kernel hilbert spaces and the randomized gacv," Technical Report 984rr, University of Wisconsin, Department of Statistics, Tech. Rep., 1998. 3

[21] L. Xu, K. Crammer, and D. Schuurmans, "Robust support vector machine training via convex outlier ablation," in *Proc. of the Twenty-First National Conference on Artificial Intelligence (AAAI)*, 2006. 3

[22] L. Wang, H. Jia, and J. Li, "Training robust support vector machine with smooth ramp loss in the primal space," *Neurocomputing*, vol. 71, pp. 3020–3025, 2008. 3

[23] X.-C. Zhou, H.-B. Shen, and J.-P. Ye, "Integrating outlier filtering in large margin training," *Journal of Zhejiang University - Science C*, vol. 12, no. 5, pp. 362–370, 2011. 3

[24] Y.-J. Lee and O. Mangasarian, "Ssvm: A smooth support vector machine for classification," *Computational Optimization and Applications*, vol. 20, no. 1, pp. 5–22, 2001. 3

[25] Z. Kou, J. Xu, X. Zhang, and L. Ji, "An improved support vector machine using class median vectors," in *Proc of 8th Intl Conf on Neural Information*, 2001. 3

[26] Y. Zhan and D. Shen, *Increasing Efficiency of SVM by Adaptively Penalizing Outliers*. Springer Berlin / Heidelberg, 2005, ch. Energy Minimization Methods in Computer Vision and Pattern Recognition, Lecture Notes in Computer Science, pp. 539–551. 3

[27] M. O.L., *Nonlinear Programming*. Society for Industrial Mathematics, 1994, vol. 10. 4

[28] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "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. 5

[29] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 57–64. 6

[30] P. Lucey, S. Lucey, and J. Cohn, "Registration Invariant Representations for Expression Detection," in *International Conference on Digital Image Computing: Techniques and Applications: Techniques and Applications*, 2010, pp. 255–261. 6