Distribution-based Iterative Pairwise Classification of Emotions in the Wild Using LGBP-TOP

Timur R. Almaev
The University of Nottingham
Mised Reality Lab
psxta4@nottingham.ac.uk

Anıl Yüce
Signal Processing Laboratory(LTS5)
École Polytechnique Fédérale de Lausanne, Switzerland
anil.yuce@epfl.ch

Alexandru Ghitulescu
The University of Nottingham
Mixed Reality Lab
psyadg@nottingham.ac.uk

Michel F. Valstar
The University of Nottingham
Mixed Reality Lab
michel.valstar@nottingham.ac.uk

ABSTRACT

Automatic facial expression analysis promises to be a game-changer in many application areas. But before this promise can be fulfilled, it has to move from the laboratory into the wild. The Emotion Recognition in the Wild challenge provides an opportunity to develop approaches in this direction. We propose a novel Distribution-based Pairwise Iterative Classification scheme, which outperforms standard multi-class classification on this challenge data. We also verify that the recently proposed dynamic appearance descriptor, Local Gabor Patterns on Three Orthogonal Planes, performs well on this real-world data, indicating that it is robust to the type of facial misalignments that can be expected in such scenarios. Finally, we provide details of ACTC, our affective computing tools on the cloud, which is a new resource for researchers in the field of affective computing.

Keywords

Emotion Recognition, Facial Expression Recognition, LGBP-TOP, Multi-class classification

1. INTRODUCTION

Analysis of facial expressions is being widely investigated in the computer vision community mainly due to the numerous types of applications that it promises to enable, such as intelligent human computer interaction, affective disorders diagnosis and treatment, national security, marketing applications and research in theoretical psychology. The most common target is to recognise the 6 (or 7, if one includes contempt) universal facial expressions which correspond to the 6 basic emotions, following the theories proposed by Ekman [9] and earlier marked by Darwin [6].

ICMI 2013 Sydney, Australia

While the field of automatic facial expression analysis has come a long way since its inception, a string of recent expression and emotion recognition challenges (e.g. [19, 16, 17]) have made clear that success rates depend for a great deal on the formulation of the task and the source of the data. The first two AVEC challenges ([17] and [16]) are also good examples of the current trend of continuous emotion recognition, implying continuity in both the affective dimensions and time. Embracing continuous affect allows any expressive behaviour to be annotated for its affective quality, whereas the six basic emotions are not frequently displayed (apart from happiness perhaps). On the other hand, the six basic expressions are universally encoded and decoded, easy to grasp by non-experts, and when displayed convey a very strong expressive message that should not be ignored. It is thus entirely valuable to pursue the recognition of basic expressions under realistic image acquisition conditions, as proposed in the Emotion Recognition in the Wild Challenge 2013 (EmotiW 2013, [7]).

In order to be able to use facial expressions of basic emotion recognition systems in real-world applications, such systems should be trained and tested on naturally generated (spontaneous) expressions recorded in real-world conditions. One of the main problems in obtaining such data is evoking meaningful instances of displays representing the proposed set of discrete emotions on the recorded subjects. Due to the difficulty of this task most of the commonly used facial expression databases, like CK+([15]), MMI([18]) or GEMEP([2]), contain subjects either posing the expressions directly as instructed, or acting them out in a manner displaying the underlying emotion using the method acting technique. Recently, this problem has been addressed by a number of researchers (e.g. [23, 14, 3]) and spontaneous expression databases have slowly started to be more widely available and more richly annotated. However, the problem of obtaining examples of these expressions in a natural environment remains, since all known databases of spontaneous expressions were recorded in controlled lab conditions.

The AFEW database, which forms the basis for the EmotiW challenge [8], aims to tackle this deficit by using extracts from movies where either one of the six basic emotions or a neutral emotion is expressed. The expressions are displayed

by professional actors from a wide age range (1-70 yr) and multiple ethnicities, in the context of the film scenes. While not being fully natural (the scenes are taken from films, after all), the expressions are more spontaneous and natural than those obtained by simply asking an actor to portray an emotion, as is the case of many other databases. But more importantly, the scenes also include various illumination modes and conditions like occlusions, extreme head pose and the presence of multiple faces, replicating real life conditions.

The scenes representing the six basic emotions are selected using the subtitles for the deaf and hearing impaired (SDH) and closed caption (CC) subtitles, extracting those scenes that contain keywords like '[CHEERS]', '[SHOUTS]', or '[SURPRISED]' in the subtitles [8]. The keywords were chosen to correspond with one of the 6 basic emotions (anger, fear, sadness, surprise, disgust and happiness) or the neutral case, automatically generating emotion label candidates for the scenes. These labels were then checked by human annotators and the videos that were found irrelevant to the generated emotion label were eliminated from the dataset.

The data is challenging, not only because of the large variety of scene conditions, but also due to the way that the sequences are labelled. There have been many debates in the community on what kind of classification the automatic expression recognition systems should focus on: the emotion felt by the person being observed, the intended message of the display, or the message as decoded and understood by the observer. In the AFEW data we are faced with another kind of classification, where captions and subtitles are used as a basis to label scenes, thus the scene context is labelled together with the emotions displayed by the actors, resulting in a compound label. Although these labels were later checked by human annotators for relevance to the labelled emotion, this kind of labelling encourages situations in which none of the conventional modalities, such as the actor's voice or face, contain significant information related to the labelled emotion. This makes the classification task more challenging compared to other facial expression databases.

It would thus make sense to employ a cinematographic approach to infer what the director of the film intended the emotion transfer to be. However, due to time constraints this paper focuses on displays of basic emotion recognition in video using facial expressions. Automatic facial expression recognition systems use either geometric (facial point) or appearance (texture) based features or both of them in combination to infer an underlying emotion or to simply detect a facial action. For an overview of the work in this area the reader is referred to the meta-analysis of the recently conducted challenge on the GEMEP-FERA database [19] or one of the recent surveys [21, 10]. More recent work on automatic facial expression recognition that addresses some of the major outstanding issues includes a Partial Least Squares approach by Güney et al. to attain pose-independent emotion recognition [11] and Chu et al., who proposed a Selective Transfer Machine to improve personalised facial expression analysis without the need to train person-specific models [5].

In this work, we propose a novel multi-class classification scheme based on iterative pairwise classification initialised by an expected class distribution to perform the classification of the sequences from the AFEW database. Classical methods of multi-class classification are unable to deal with the complicated nature of this data. The general idea behind our proposed Distribution-based Iterative Pairwise Classification (DIP-C) is that the expected distribution of classes can be approximated *a priori*. DIP-C uses this distribution to randomly initialise the set of test labels with a label assignment based on it and iteratively updates the predicted labels by comparing pairs of test instances, only adjusting the predicted labels when significant evidence is found from the binary classifiers based on the pair of examples.

We also evaluate the recently proposed dynamic appearance descriptor Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP). Static appearance descriptors such as Local Binary Patterns (LBP) and variants such as Local Phase Quantisation ([12]), Local Gabor Binary Patterns ([22]) or Multiple Local Curvature Gabor Binary Patterns [20] have been successfully used for both facial expression and action unit detection. More recently, it has been shown that the dynamic extension of LGBP, LGBP-TOP, outperforms similar appearance based features in Action Unit detection, and some evidence was provided suggesting that contrary to expectations TOP features are more robust to face alignment errors than their static counterparts, which would make them more suitable for expression recognition in the wild [1].

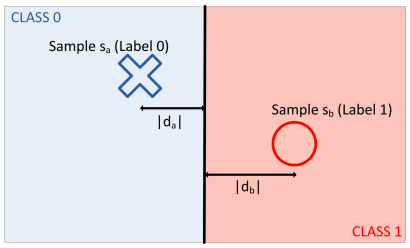
The main contribution of this work is the novel multi-class classification method (DIP-C) and the use of LGBP-TOP features with the complex AFEW database, which are shown to significantly outperform the challenge baseline results in emotion recognition accuracy.

The rest of the paper is organised as follows: In Section 2 we explain the SVM based iterative classification scheme, in Section 3 we describe the dynamic appearance descriptor used for classification. Section 4 presents our results on the development and test portions of the AFEW database as well as comparison with other methods. Finally, we conclude the paper with a brief discussion in Section 6.

2. DISTRIBUTION-BASED ITERATIVE PAIRWISE CLASSIFICATION

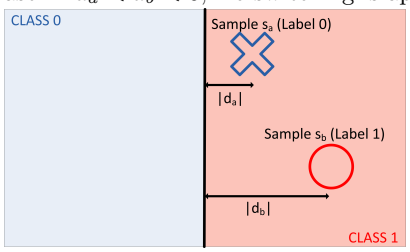
The common solution to multi-class problems with low number of training examples, like basic emotion classification, is to train many one-vs-one or one-vs-all binary classifiers and use a voting strategy respectively maximum decision function value determination to decide on the final classification. This is mainly because classifiers which are multi-class by nature (e.g. random forests) generally require a lot of training data to learn the decision function and optimise their many parameters. Due to the challenging character of the data in question, conventional methods for multi-class classification are not sufficient for solving this particular multi-class problem, as they cannot be guaranteed to adhere to the *a priori* known approximate class distribution of a set of test examples. Therefore, in this work we propose a framework that assigns an initial label to each sample based on this distribution and fine-tunes the label guessed for pairs of examples using binary SVMs trained in a one-vs-one fashion. For this task we have trained 21 one-vs-one linear SVMs, each optimised for the cost parameter C on the validation set, using the publicly available LibSVM implementation [4].

We start the classification process by assigning a class label to each example in the test set in a random fashion based on the distribution of classes in the data used to train the

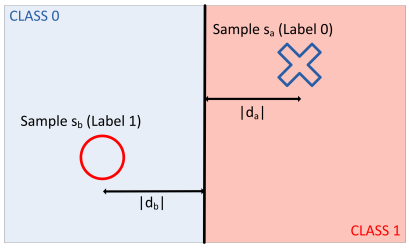


(a) Case 1: $d_a < 0 < d_b$; No switching is applied

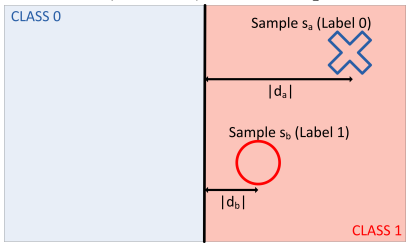
(b) Case 2: $d_a < d_b < 0$; No switching is applied



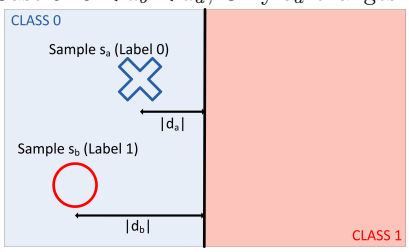
(c) Case 3: $0 < d_a < d_b$; No switching is applied



(d) Case 4: $d_b < 0$, $d_a > 0$; Both samples switch labels



(e) Case 5: $0 < d_b < d_a$; Only s_a changes label



(f) Case 6: $d_b < d_a < 0$; Only s_b changes label

Figure 1: Possible one-vs.-one classification results and corresponding switching operations applied

binary classifiers. In our case we have 7 classes, and for this initialisation the probability of each class being assigned is proportional to the number of examples belonging to that class in the training and development partitions. The idea behind this initialisation is the assumption that a training set of data will possess a similar distribution of classes compared to the test sample. We also show in the results section (Sec. 4) that this initial distribution does not affect the final classification accuracy significantly, as long as a reasonable assumption is made (e.g. uniform distribution of classes).

Next, we iteratively improve our prediction. In each iteration we pair each example in our test set with another example randomly chosen from the same set and apply the one-vs-one linear SVM corresponding to the initially guessed labels for these 2 examples; e.g. if the picked examples are from the anger and fear classes respectively, we apply the classifier anger vs. fear to both of the examples. Then, depending on the distances of each example to the decision hyperplane of the corresponding classifier we apply a label update operation. To explain this operation let us assume a case where the first sample s_a of guessed class L_0 is paired with sample s_b of guessed class L_1, with distances to the decision hyperplane of classifier $0vs1$ denoted by d_a and d_b, respectively. Let us also affirm that the classifier would normally assign label L_0 if this distance is smaller than 0.

We now define 4 tests to determine whether labels should be updated (illustrated in Figure 1). The tests are only subtly different from ordinary binary classification but, as experiments will show, result in a significant classification improvement. The first test is whether $d_a < d_b$, which means both examples are closer to their assigned classes than the example they are compared to, so we keep their assigned labels. This covers all three cases where $d_a < 0 < d_b$, $d_a < d_b < 0$ and $0 < d_a < d_b$, as represented in Figures 1a, 1b and 1c respectively. The second case is where $d_a > 0$ and $d_b < 0$ (Fig. 1d). This means that according to the classifier both examples are incorrectly labelled and each would be better represented by the class that they are compared to. So, we switch the labels of both examples, assigning s_a the label L_1 and s_b the label L_0. The third test is $d_a > d_b > 0$ (Fig. 1e), where we switch the label of s_a to L_1 but keep s_b's assigned label, which is also L_1. This is because compared to s_b, s_a is closer to class L_1, yet we do not have significant evidence for sample s_b to change its label to L_0. The last test is whether $0 > d_a > d_b$ (Fig. 1f) for which we switch the label of s_b and keep that of s_a, for the same reasoning as in the third case.

The main idea behind this label fine-tuning is to collect evidence for each example using other examples and to maintain the initial distribution of classes unless an indicative evidence is present. Another way of viewing this process is that normal classification is carried out on a pair of examples except when the classifier is less confident for an example compared to the other half of the pair which already belongs to a correct class, as in Figures 1b and 1c.

For every iteration this process is applied to all examples, and the process is repeated N times. N is determined empirically by testing on the validation set, details of which are given in Section 4. At each iteration the only restriction in pairing examples is that they do not belong to the same

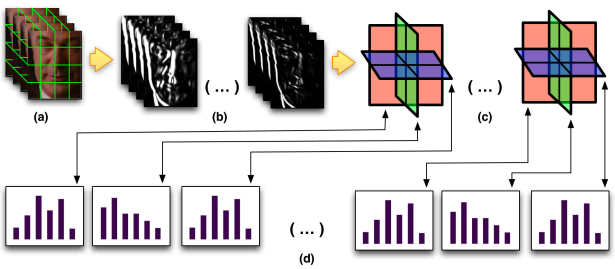


Figure 2: Extraction procedure for Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP).

class. Ideally, each example is compared to many other examples, and is inspected by each of the 21 different classifiers many times. This scheme provides an accumulation of confidence with regards to belonging to a class for each example and as shown in the following section results in a significant increase in accuracy compared to conventional multi-class SVM classification methods. Here, we have chosen SVM as the one-vs-one classifier, but the method is a general one and applicable to any binary classifier that outputs an ordinal value proportional to the confidence or probability of classification.

3. LGBP FROM THREE ORTHOGONAL PLANES

The approach we have chosen for visual feature extraction is the recently published Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [1] dynamic appearance descriptor, combined with Principal Component Analysis (PCA) for dimensionality reduction. Since the focus of this paper is to establish the merits of LGBP-TOP and DIP-C we did not use any additional face/head detection nor face alignment or registration. Instead we utilise the aligned faces provided by the challenge obtained using the mixture of parts based detector [24], which has proven to be quite robust to conditions such as extreme head-poses and facial expressions. This should allow for a better comparison with the baseline results.

Constructing LBGP-TOP features consists of applying Gabor filters of various frequencies and orientations on the frames of a video sequence prior to applying the LBP transform and computing histograms over fixed spatio-temporal windows, finally using the histogram bin values as image sequence descriptors. More specifically, on each input frame we apply 18 Gabor filters of 3 spatial frequencies ϕ (Eq. 1) and 6 orientations (Eq. 2).

$$\phi = \left(\frac{\pi}{2}, \frac{\pi}{4}, \frac{\pi}{8}\right) \tag{1}$$

$$\theta = \left(\frac{k\pi}{6}, k \in \{0...5\}\right) \tag{2}$$

We apply the uniform LBP transform to the x, y, x, t, and y, t intersections of sequences of Gabor filtered images divided into 4 by 4 spatio-temporal cubes (see Fig. 2), where the temporal window is fixed to 5 frames. Histograms are computed for the three intersection orientations for each cube and then concatenated. This way we obtain for each frame $18 \times 4 \times 4 \times 59 \times 3 = 50\,976$ features. Since the goal of the EmotiW challenge is to classify video sequences instead of single frames, we compute the mean LGBP-TOP features over all frames of a sequence. Computing the mean over frames also serves to decrease the possible noise caused by face detection or registration errors, including the few cases where an additional face other than the one in most of the frames of the sequence appears in the aligned faces sequence provided by the challenge.

One of the biggest disadvantages of LGBP-TOP is the very large number of features generated per block of frames, which is equal to 50 976 for the selected set of Gabor and temporal LBP configuration. Thus, we apply dimensionality reduction using PCA, which provides the transformation of the feature set into a lower dimensional space composed of components giving the largest variance among a set of examples, which is in our case the training partition of the dataset. Although PCA does not take into consideration the final classification task when reducing the dimension, it helps to significantly reduce the redundancy and noise among the features, which in our case are the bins of histograms generated. After the PCA, the size of the feature set reduces to 378 per sequence.

LGBP-TOP, which combine motion and appearance features extraction with a prior Gabor wavelet filtering, has recently shown to outperform similar texture descriptors such as LBP,LBP-TOP and LGBP [1]. An advantage specifically worth underlining is that it is more robust against facial misalignment compared to its static counterpart LGBP. Due to the wild nature of the data used in this challenge, this robustness property is particularly useful for the overall recognition accuracy of the system. We compare the accuracy of LGBP-TOP features with LGBP as well as LBP-TOP features, which are used to obtain the baseline results, in the results section (Sec. 4).

4. EXPERIMENTAL RESULTS

This section provides the results obtained on the test and development partitions for the EmotiW challenge using the proposed feature extraction and classification method as well as a comparison with various other techniques.

The first results we present are for the experiments we have performed to determine the type of kernel to be used for the SVM classifiers. For this purpose we have trained linear, histogram intersection (using the implementation by Maji et al. [13]) and RBF kernel SVMs on the training partition of the dataset and tested on the development partition. In all cases SVM parameter optimisation was performed, which includes search for the best C value for linear and intersection kernels with additional gamma value optimisation for the RBF kernel. Figure 3 shows the results obtained using LGBP and LGBP-TOP features using these 3 types of kernels in the conventional multi-class SVM classification using one-vs-one classification scheme. Shown are the overall accuracies obtained for the 7 classes (angry, disgust, fear, happy, sad, surprise and neutral) on the development (a.k.a. validation) partition.

One of the main observations from Figure 3 is the superiority of LGBP-TOP features over the static LGBP using all three types of kernels, as expected. The best performance is obtained with the RBF kernels, but since it is not significantly superior to the linear kernel we apply Occam's razor

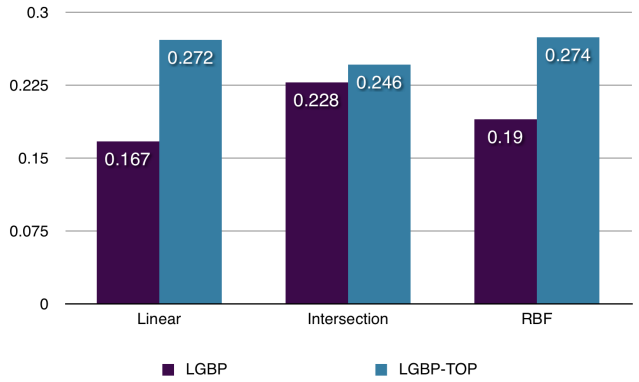


Figure 3: Comparison of LGBP and LGBP-TOP features with different kernels used for one-vs-one multiclass SVMs tested on the development set

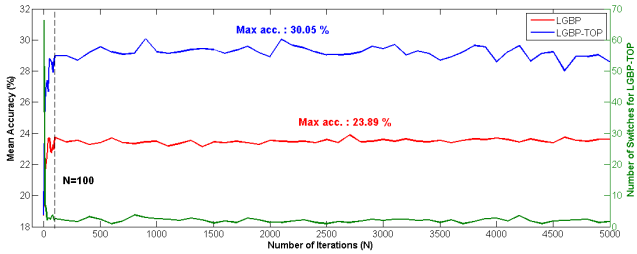
and prefer to use the linear kernels in further tests, since they are known to generalise better on unseen data.

We apply the iterative one-vs-one classification scheme (DIP-C) as explained in detail in Section 2 using both LBGP-TOP and LGBP features. This method applies one of the 21 one-vs-one classifiers on each sample multiple times iteratively, and the choice of the classifier is made depending on the currently assigned pair of labels. Since the pairing is a stochastic process, it is expected that every run gives a slightly different result, which also depends on the number of iterations over the dataset. To observe the dependency on the number of iterations we have measured the accuracy after different numbers of iterations (0 to 5000) for both kinds of features (averaged over 5 repeated trials). The results of this experiment are shown in Figure 4. We also show in the same plot how the number of label switches that were executed evolves as the number of iterations grows.

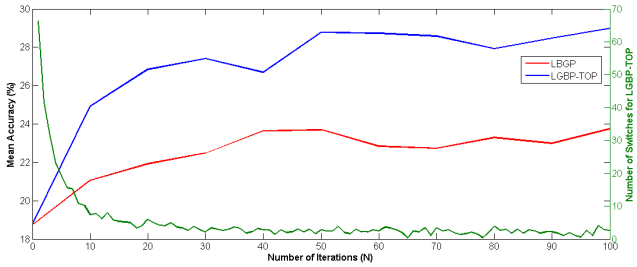
As can be observed from Fig. 4, there is no clear optimal value for the number of iterations for either of the descriptors. However, there is a clear threshold at roughly 100 iterations, after which the accuracy does not change significantly. There is also a consistent difference in recognition performance between LGBP and LGBP-TOP regardless of the number of iterations, once again marking the superiority of the chosen features. The best average accuracy obtained is 30.05% with a standard deviation of ±1.14%.

While we cannot prove the convergence of DIP-C, the green line in Fig. 4 shows the number of label switches applied at each iteration, demonstrating clearly that the algorithm stabilises after a burn-in time, after which only a few switches are made at each subsequent iteration. Unfortunately, this burn-in time is shorter than the 100 iterations required to obtain maximum accuracy, which means we cannot use the number of switches directly as the cut-off point for training.

Another important component of the DIP-C algorithm is the initialisation phase. For the definitive challenge results on the testing partition, we have used the distribution of classes in the training partition to perform the initialisation, with the assumption that, in general, the set of training examples can approximately represent the distribution of unseen data over the different classes. To demonstrate the effect of this initialisation, we have performed an additional set of experiments on the development partition. In particular, we have performed 5 trials for three different initialisations; using i) the training set distribution, ii) a uniform distribution and iii) the actual development set distribution, results of which can be seen in Table 1.



(a) From 0 to 5000 iterations



(b) Detailed plot for 0 to 100 iterations

Figure 4: Number of iterations vs. mean accuracy over 5 trials for LGBP and LGBP-TOP features tested on the development partition, in addition to the average number of label switches performed at each iteration for the LGBP-TOP case over 5 trials (in green).

Table 1: Mean accuracies and standard deviations over 5 trials using different initial class label distributions.

Initialization	Mean acc. (%)
Training Set	30.05 ± 1.14
Uniform	29.14 ± 1.02
Development Set	29.34 ± 1.27

We observe from Table 1 that there is very little difference in accuracy between the 3 initialisation distributions and that the proposed method is robust against the choice of initial class distribution, as long as a reasonable assumption is made, such as using a training set of examples or a uniform distribution.

Finally, we have tested our method giving the best accuracy, which is the LGBP-TOP features combined with the DIP-C iterative classification scheme, on the testing partition of the challenge dataset. The results obtained are presented in Figure 5 along with the corresponding results on the development partition for the purpose of comparison with the challenge baseline method and simple multiclass SVM classification using the same features.

Clearly, the proposed approach (DIP-C) outperforms both the baseline method provided and the straightforward multi-

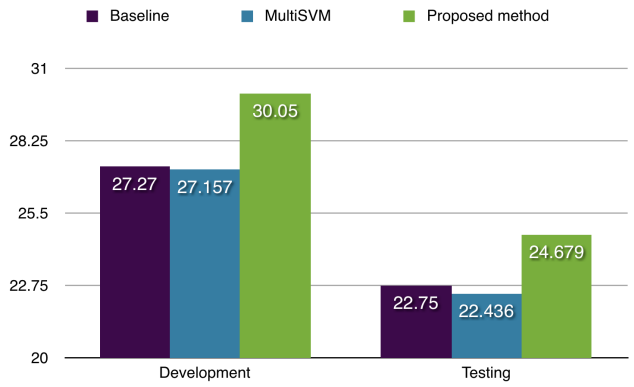


Figure 5: Accuracy results of the one-vs.-all Multiclass SVM and our proposed method on the development and testing partitions compared to the baseline results.

Table 2: Confusion Matrix on the AFEW test partition for the six basic expressions plus neutral.

P \ T	Ang	Disg	Fear	Hap	Neut	Sad	Surp	Total
Ang	**25**	2	5	3	4	2	13	54
Disg	11	**3**	3	8	4	5	15	49
Fear	9	0	**4**	3	10	1	6	33
Hap	7	2	1	**27**	1	6	6	50
Neut	11	1	7	6	**8**	6	9	48
Sad	7	3	4	13	2	**6**	8	43
Surp	7	3	5	4	9	3	**4**	35
Total	77	14	29	64	38	29	61	

class SVM approach. Although we obtain better accuracy than the baseline method the classification results are not necessarily successful, as can be observed from Fig. 5 and the confusion matrix in Table 2. Apart from the most obvious reason that it is very hard to label "wild" data in general, we suggest that another important factor causing the low accuracy is how the data is labelled (summarised in Section 1). This kind of labelling from captions of movie scenes does not guarantee a particular emotional state expressed on a face, but implies in the best case the emotion expressed by the whole scene, of which a face is just a small part. A possible method we propose to overcome this complex classification task is to use additional modalities in combination with the facial information, such as audio, colour, motion and others that might give information about the scene context.

5. AFFECTIVE COMPUTING TOOLS ON THE CLOUD

With the field of facial expression analysis and affective computing in general maturing, there is a new drive for authors to make their proposed systems available to other researchers, be it to allow comparison with future approaches or to be used as a data analysis tool (e.g. by psychologists or researchers working on social robot interaction). Traditionally this need would be addressed by making source or binary code available through the authors' websites. However, this invariably leads to problems with variations in end users systems setup (e.g. different operating systems or installed libraries), problems in authoring bug fixes or improvements to existing code and consequently deploying them to the end users, and it often assumes a level of technical knowledge that not all interested parties may possess.

We therefore created a new cloud-based solution to sharing face analysis technology, called Affective Computing Tools on the Cloud, or ACTC. This online resource simplifies the process of video analysis for the researcher.

The process consists of the following steps:

1. Register an account for your research group
2. Upload the videos/images
3. Choose a set of tools to apply to the images/videos
4. Download the archive of results

A number of affective computing tools made by our lab will be available through this tool, including the work presented here. ACTC uses user-level access control and users will be divided into two categories: team lead and team member. A designated team lead user is responsible for a research team, and has the ability to create virtually any number of normal team member users. There will be a limit on the disk quota and the number of parallel jobs that can be ran at any one time by the research team, but the team lead can set targets by prioritising, cancelling or deleting jobs from their own users. A normal user is limited to one job at a time and has full control only over their own jobs. We hope that this gives the flexibility and control required in an academic research environment.

The videos are organised into folders and as soon as a processing tool is selected they become jobs and they are added to a queue. The only limit on the length of the queue is the disk quota associated with the user. ACTC runs on a parallel computing cluster that makes heavy use of Graphics Processing Units (GPUs). The resource is freely available for academic research upon registration online at http://actc.cs.nott.ac.uk.

6. CONCLUSION

In this paper a comparison between a number of facial expression recognition techniques has been presented in the context of the EmotiW2013 challenge dataset. In particular, the recently introduced LGBP-TOP appearance descriptor is compared against its more traditional static counterpart LGBP in order to show how well LGBP-TOP, previously shown to perform very well on laboratory recorded data, deals with challenging wild data. The results obtained clearly show that LGBP-TOP gives a significant improvement over LGBP in all cases studied. In addition, we have introduced a specialised multi-class classification approach called Distribution-based Iterative Pair-wise Classification (DIP-C), based on iterative one-versus-one classification of randomly picked pairs of instances with randomly assigned labels. DIP-C is meant to be used in situations where predictions have to be made on a set of test examples and a reasonable expectation of the posterior class distribution is known *a priori*. Experimental results show that this classification scheme significantly improves overall recognition performance of the system on both given development and testing datasets compared to the baseline result and the simple multiclass SVM.

The reason why the overall recognition accuracy is quite low despite all the advances introduced, lies, in our opinion,

in the way the challenge data was labelled. Since the data was labelled according to subtitles, it is clearly limiting to solely use facial expression recognition to classify this particular data, whose labelling depends on the context of the scene and many other factors. Therefore we propose for this particular dataset, as a future work to improve recognition accuracy, to move from facial analysis to a whole scene analysis and determining which factors are really related to the emotion labels.

7. REFERENCES

[1] T. Almaev and M. Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Proc. Affective Computing and Intelligent Interaction*, 2013.

[2] T. Bänziger and K. R. Scherer. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook*, pages 271–294, 2010.

[3] C.Busso, M.Bulut, C.Lee, A.Kazemzadeh, E.Mower, S.Kim, J. Chang, S. Lee, and S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335 – 359, 2008.

[4] C. C. Chang and C. J. Lin. Libsvm: A library for support vector machines. *Intelligent Systems and Technology, ACM Transactions on*, 2(3):1–27, 2011.

[5] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[6] C. Darwin. *The expression of the emotions in man and animals*. Oxford University Press, 1998.

[7] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proc. ACM Int'l Conf. Multimodal Interaction*, 2013. in print.

[8] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):0034, 2012.

[9] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[10] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259 – 275, 2003.

[11] F. Güney, N. Arar, M. Fischer, and H. Ekenel. Cross-pose facial expression recognition. In *International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, 2013.

[12] B. Jiang, M. F. 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*, pages 314–321. IEEE, 2011.

[13] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[14] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5–17, 2012.

[15] P. P. Lucey, J. F. 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. *Computer Vision and Pattern Recognition Workshops (CVPRW),IEEE International Conference on*, pages 94–101, 2010.

[16] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM, 2012.

[17] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011–the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011.

[18] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. Intl. Conf. Language Resources and Evaluation, Workshop on EMOTION*, pages 65–70, 2010.

[19] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta analysis of the first facial expression recognition challenge. *Systems, Man, Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(4):966–979, 2012.

[20] A. Yuce, N. M. Arar, and J.-P. Thiran. Multiple local curvature gabor binary patterns for facial action recognition. In *4th International Workshop on Human Behavior Understanding, in conjunction with ACM Multimedia*, 2013.

[21] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39 – 58, 2009.

[22] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition. *Computer Vision, 10th IEEE International Conference on*, 1:786–791, 2005.

[23] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *Proceedings of 10th IEEE International conference on automatic face and gesture recognition (FG'13)*, 2013.

[24] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.