Why is Facial Expression Analysis in the Wild Challenging?

Tobias Gehrig
Institute for Anthropomatics
Karlsruhe Institute of Technology, Germany
tobias.gehrig@kit.edu

Hazım Kemal Ekenel
Faculty of Computer and Informatics
Istanbul Technical University, Turkey
Institute for Anthropomatics
Karlsruhe Institute of Technology, Germany
ekenel@itu.edu.tr

ABSTRACT

In this paper, we discuss the challenges for facial expression analysis in the wild. We studied the problems exemplarily on the Emotion Recognition in the Wild Challenge 2013 [3] dataset. We performed extensive experiments on this dataset comparing different approaches for face alignment, face representation, and classification, as well as human performance. It turns out that under close-to-real conditions, especially with co-occurring speech, it is hard even for humans to assign emotion labels to clips when only taking video into account. Our experiments on automatic emotion classification achieved at best a correct classification rate of 29.81% on the test set using Gabor features and linear support vector machines, which were trained on web images. This result is 7.06% better than the official baseline, which additionally incorporates time information.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications; I.4.m [**IMAGE PROCESSING AND COMPUTER VISION**]: Miscellaneous

Keywords

Facial Expression; Emotion; EmotiW; DCT; LBP; Gabor; FACS; SVM

1. INTRODUCTION

Facial expression analysis is a popular topic, which attracts more and more interest. Research on it already goes back to more than two decades [7, 13, 20, 23]. However, a lot of studies are still using posed or at least lab data for experiments. One of the main reasons for this is that it is hard and time consuming to collect and annotate realistic data. But for real applications one needs a system which can cope with spontaneous expressions, which involve different facial muscles and show different dynamics compared to posed expressions [1]. One step towards more realistic data and evaluations was taken by the *FG 2011 Facial Expression Recognition and Analysis Challenge* (FERA2011) [21, 22]. Nevertheless, the data used in that challenge was still collected in a lab setting with relatively frontal faces, since the actors interacted towards the camera. Recently, the *Acted Facial Expression in the Wild* (AFEW) dataset [4] has been published, which targets to fill this gap in the datasets. It is a collection of clips from movies with labels of seven basic emotions (anger, disgust, fear, happy, neutral, sad, surprise). As such, it resembles close-to-real conditions in terms of emotional colored sequences. Even though the name of the database suggests that it is focusing on facial expressions, the clips also contain audio tracks and due to the setting also visual context knowledge, which aids in the decision of the appropriate emotion. This dataset contains realistic challenges like different illumination conditions, occlusion, pose variations, relative spontaneous emotional expressions. Subsets of the AFEW dataset are used in the *Emotion Recognition in the Wild Challenge* (EmotiW2013) [3] to provide a common benchmark for comparison of approaches targeting facial expression analysis in the wild.

In this study, we analyze a diverse set of face alignment methods, face representations, and classifiers for their usability in emotion classification on close-to-real data. In terms of face alignment, we compare eye-based alignment to fiducial point warping based on a *mixture-of-parts* (MoP) model. The face representations we looked at are block-based *discrete cosine transform* (DCT) [5], *local binary patterns* (LBP) [12], *Gabor*, *Gabor DCT* (GDCT), and facial *action unit* (AU) intensities [6]. Finally, for classifying the emotions we selected a *nearest neighbor* (NN) based classifier, a *nearest mean* (NM) based classifier, and 1-versus-1 multi-class *support vector machines* (SVMs) with linear, second-order polynomial, and RBF kernels. Additionally, we did human evaluations to determine the difficulty of the task for humans, when they use only the video track to base their decision on.

The content of the paper is organized as follows. We will present the methods we selected for the different stages of automatic facial expression analysis, i.e. face alignment, face representations, and finally the classifiers in Section 2. The used dataset will be described in Section 3, followed by the experimental section in Section 4, where we studied the influence of face alignment, face representation, head orientation, classifiers, and training data, as well as human performance and how the automatic approaches perform on

EmotiW'13, December 9, 2013, Sydney, NSW, Australia

http://dx.doi.org/10.1145/2531923.2531924.

a human agreed subset. Finally, in Section 5 we will give conclusions and suggestions for possibilities about how to improve evaluations on such realistic data.

2. METHODOLOGY

The general approach to automatic facial expression analysis consists of multiple stages in the processing. First a face is detected and aligned to some reference, then a compact face representation is derived, and finally the corresponding emotion is estimated using a classifier. In following subsections, we describe the methods we selected in this work for comparison within the individual stages.

2.1 Face Alignment

We investigated two alignment methods, a simple eye-based alignment [8] and a fiducial points based warping [24], as provided by the organizers of the EmotiW2013 challenge.

The process for the eye-based alignment is as described by Gehrig and Ekenel in [8]. First, a *modified census transform* (MCT)-based face detector is used to detect faces [10]. Since that detector is trained on frontal faces, this already reduces the number of frames to the ones showing almost frontal faces. The next step is detecting eyes using a MCT-based eye detector. The processing continues only if both eyes were found. Using these eye locations the face image is scaled and rotated such that the eyes are always located at a predefined position in the aligned face, i.e. with a fixed interocular distance, on a specific row of the cropped image. For the fiducial points based alignment, the aligned face images provided by the EmotiW 2013 Challenge organizers are used. These are based on a *mixture-of-parts* (MoP) based face and fiducial points detector [24]. For the alignment based on the detected fiducial points, an affine transform, followed by warping, using the Matlab functions `cp2tform` and `imtransform`, was applied.

2.2 Face Representation

We compare five different kind of face representations: block-based *discrete cosine transform* (DCT) [5], *local binary patterns* (LBP) [12], *Gabor*, *Gabor DCT* (GDCT), and facial *action unit* (AU) intensities [6].

For the block-based DCT, we do the processing as proposed by Gehrig and Ekenel [8]. First, the aligned face image is divided into non-overlapping blocks of $N \times N$ pixels. Then, each block is transformed using a two-dimensional type-II DCT. The DCT coefficients are extracted using zig-zag scanning and only the first few are kept for further processing. Finally, the block coefficients are normalized by dividing each coefficient by its standard deviation, followed by a normalization of the resulting block feature vector to unit norm. The feature vector for the whole face is formed by concatenating all the block feature vectors.

When using the LBP-based face representation, the aligned image is first transformed using a uniform LBP operator. Then, the transformed image is split into non-overlapping blocks of size $N \times N$. Finally, a histogram over the distribution of the LBP codes in each block is calculated and appended to the final feature vector, similar to the one in Shan et al. [17].

The face representation using Gabor filters is performed similarly to the one proposed by Richter et al. [14]. First, the aligned image is filtered using a bank of Gabor filter with n orientations and m scales of which only the magnitude is used in the further processing. Then, the individual magnitude responses are downscaled by a specific factor a to decrease the feature dimensionality. Finally, all the downscaled responses are concatenated to one big feature vector.

The GDCT based face representation is a novel approach. It is basically consecutively applying the Gabor face representation without the downscaling and then the block-based DCT. This makes it possible to use the full sized Gabor magnitude responses, but at the same time compressing them.

The AU intensity features are based on the definitions in the *facial action coding system* (FACS) proposed by Ekman and Friesen [6] and allow an objective description of facial expressions closely related to facial muscle activations. Here, we estimate them using an approach based on the AU detection framework proposed by Gehrig and Ekenel in [9]. This framework uses *partial least squares* (PLS) to detect the activation of different AUs. We extended that to AU intensity estimation by using the FACS intensity labels instead of the binary activation labels to train a model. This model was then used to determine the AU intensities for the individual frames. Using AU intensities for the estimation of the emotion of less constrained sequences was already proposed by Littlewort et al. [11], but they used a more complex feature vector containing multiple statistical moments over time to incorporate dynamic information as well as head pose related information.

2.3 Classification

In this work, we compared several approaches for the classification of the seven basic facial expressions from emotions (angry, disgust, fear, happy, neutral, sad, and surprise). We selected a *nearest neighbor* (NN) classifier, a *nearest mean* (NM) based classifier, and 1-versus-1 SVMs with linear, second-order polynomial, and RBF kernels. Before the actual training or classification takes place for the NN and SVM classifiers the feature vectors are normalized to be zero-mean and have unit variance.

The NN classifier simply uses the L2 distance between the features of the test samples and the ones from the training set to determine the emotion for a given test frame. To determine the emotion for a whole clip the distances to the nearest neighbor per class are accumulated over all frames of a clip and the class with the lowest sum is chosen as the estimate for that clip.

In case of the NM based classifier, for each class the mean and the covariance of the feature vectors over all frames of the sequences belonging to that class is calculated. On testing, the feature vectors over all frames of a clip are averaged. This clip average feature vector is then used to calculate the Mahalanobis distance to all the classes. And again, the emotion for which this distance is the smallest is selected as the estimate for that clip.

For the SVM, we used the 1-versus-1 multi-class implementation provided by LIBSVM [2]. The classifiers were trained on all frames of a sequence of the corresponding class. Similar to the NN based classifier, on estimating the emotion of a clip, the confidences per class are accumulated over all frames of a sequence and the class with the maximum accumulated confidence is selected as the estimate for that clip. The confidence of a class is determined as the normalized number of 1-versus-1 classifiers, which vote for that class.

Table 1: Statistics of the training partition of EmotiW2013

class	Subjects	# clips				# frames	
		overall	with eyes	frontal	non-frontal	overall	with eyes
Angry	35	58	55	19	39	3379	1740
Disgust	27	40	38	28	12	2699	1945
Fear	20	50	45	26	24	2838	1645
Happy	41	65	62	41	24	3807	2673
Neutral	29	63	60	41	22	3910	2530
Sad	28	52	50	18	34	3173	2091
Surprise	26	52	49	29	23	2876	1542
Total	**99**	**380**	**359**	**202**	**178**	**22682**	**14166**

Table 2: Statistics of the validation partition of EmotiW2013

class	Subjects	# clips				# frames	
		overall	with eyes	frontal	non-frontal	overall	with eyes
Angry	36	59	54	22	37	3614	2048
Disgust	33	50	50	19	31	3460	2479
Fear	38	54	51	22	32	2661	1549
Happy	33	62	60	35	27	3432	2189
Neutral	30	55	51	33	22	3055	2255
Sad	32	64	56	19	45	3729	2219
Surprise	35	52	50	20	32	2731	1725
Total	**126**	**396**	**372**	**170**	**226**	**22682**	**14464**

3. DATASET

In this work, we use the *Emotion Recognition in the Wild Challenge* (EmotiW2013) dataset [3]. This dataset is based on the previously released *Acted Facial Expression in the Wild* (AFEW) dataset [4]. This dataset is a collection of clips selected from 75 movies. It was collected using a semi-automatic process, in which first the subtitles were extracted from the DVDs as well as downloaded from the internet. These were then searched for specific keywords related to the 7 basic emotions (angry, disgust, fear, happy, neutral, sad, and surprise). The corresponding clips were then extracted based on the timing information of the corresponding subtitle, and the human annotator labeled the subject and expression displayed in the clip. These clips were then filtered for those where only a single subject occurs. This resulted in 1832 video clips for the AFEW 3.0 dataset of which 1088 were used for the EmotiW2013 challenge. For evaluation purposes, this set was split into training (Train), validation (Val), and testing (Test) sets containing 380, 396, and 312 clips, respectively. More detailed statistics about the Train and Val sets are presented in Table 1 and Table 2. The tables include the numbers for the case where only the part is used for which eyes were detected.

4. EXPERIMENTS

In the following, we present the experiments we performed on the EmotiW2013 dataset.

4.1 Influence of Face Alignment

The first experiment consists of comparing the influence of the face alignment method on the correct classification rate. As stated in Section 2.1, we compare here a simple eye-based alignment [8] and a fiducial points based warping [24]. For the eye-based alignment, we use an eye distance of 31 pixels and the 26th row for positioning the eyes in 64×80 cropouts, and the appropriate multiples (1.5× and 2× of the eye distance and the eye row) for 96×120 and 128×160 cropouts. For the fiducial points based warping, we downscaled the provided aligned images to 64×80. We compared the two approaches using a 1-versus-1 SVM for the classification. The features were either extracted using block-based DCT features or Gabor magnitude filters. For the block-based DCT, we used non-overlapping blocks of size 8×8 and kept the first 10 coefficients for each block. For the Gabor filters, we used 8 scales, 5 orientations and downscaled the filtered images by a factor of 8.

The results for the comparison of the two face alignment approaches using different kind of features is presented in Table 3. When using DCT, using the MoP alignment is 1.26% better than the eye-based alignment. But comparing the two approaches directly is not really fair, since the MoP based alignment has almost for all frames of the clips aligned faces, while the eye-based alignment is only available for almost half of the frames. Thus, the classifier has more samples available in the case of MoP and might therefore generalize better. For that reason, we also compared using the MoP alignment for only those frames for which there is also an eye-based alignment available. In this case, both alignments perform equally well, when looking at the overall classification rate. When looking at the individual performances, it seems like those emotions, which are mostly visible through changes in the mouth region, i.e. disgust, happy, and surprise, are better estimated using MoP, because the mouth region is not at all directly aligned using the eye-based alignment. The average per class classification rate is slightly better here for MoP (19.46% for eye vs. 19.34% for MoP). Strangely, when using Gabor features the performance drops by 3.33% for the MoP based alignment. Even when using the original size of the provided MoP aligned face images with size 143×181, instead of 64×80, the performance stays below that of the eye-based alignment.

To get a glimpse of the goodness of the alignment, we calculated the mean faces over the sequences for both alignments. Some examples are shown in Figure 1. Here, we also see that the overall mouth region seems to be less blurry in case of MoP (reduced). One can see that the mean faces for the MoP alignment are more blurry, which suggests that also the alignment for MoP is noisy. Actually, another reason might be that MoP also detected sometimes non-faces or other faces, which might add to the blur. Thus, if we go back to the results in Table 3, it seems like DCT is better capable of coping with alignment noise than Gabor, since for DCT, eye-based alignment and MoP lead to more or less the same overall results. But Gabor gains quite a lot in terms of performance by using the eye-based alignment. So in this case the overall alignment wins over the local facial parts alignment.

4.2 Influence of Face Representation

We already saw in the previous experiment that using Gabor features improves the overall classification performance.

Figure 1: Mean faces for some clips from the validation set of EmotiW2013 using eye-based affine alignment or fiducial point based warping. The clips are from the movies "Remember Me", "21", "Friends With Benefit", and "Hall Pass", respectively.

Thus, we investigate here also other face representations to see how well they perform on the wild data. For that we chose DCT, LBP, Gabor and GDCT. For DCT and Gabor, we used the same settings as in the previous subsection. For LBP, we used a radius of 2 pixels and 8 neighboring pixels per uniform LBP operator. We extracted blocks of 16×20, 24×30, or 32×40 pixels depending on the face resolution, such that the blocks are arranged in a 4×4 grid. For the bank of Gabor filters in the GDCT face representation, we also used 8 orientations and 5 scales, but for the DCT part, we skipped the first coefficient of each block and only used the following 5 coefficients. We did not normalize the coefficients per block using the standard deviation.

The result of the comparison is presented in Table 4. Again, we observe that Gabor features outperform the other face representations. LBP and GDCT are in the middle field. Using a face resolution of 96×120 improves the performance over 64×80 and 128×160 for Gabor, LBP and DCT.

4.3 Influence of frontal vs. non-frontal faces

To determine how well a specific face alignment method and face representation works across poses, we evaluated the results on just the subset of the validation set with frontal faces, respectively non-frontal faces. The results on 64×80 faces are shown in Table 5. As expected, the overall results for the frontal subset are generally better than those for the non-frontal subset. But for classes like angry and sad, for which there are more non-frontal clips available in the training set, the performance is almost consistently better for non-frontal clips. When comparing again the face alignment methods, it seems that using DCT, it does not matter much on average which pose the test faces have as long as they are detectable. But here, the fiducial point warping (MoP-reduced) works much better on frontal clips. In case of LBP and Gabor, the performance drops quite a lot from frontal to non-frontal when using the eye-based alignment. For Gabor, the performance drop is not as huge for MoP-reduced compared to the eye-based alignment. But on the other hand, it does not reach the performance for frontal faces as with the eye-based alignment. This is also visible in the example mean faces in Figure 1, since for frontal faces, it is more blurry compared to the eye-based alignment, while

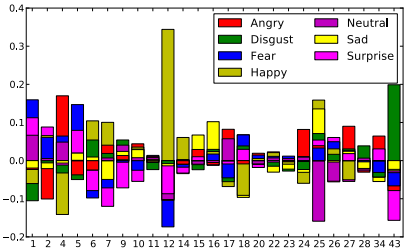


Figure 2: Overlayed plot of mean AU intensity estimates per class on the training set of EmotiW2013 normalized by the overall mean.

it is the other way around for the non-profile face. This is also due to the fact that the eye detector is trained solely on frontal eyes and thus begins to jitter as the pose changes from frontal to non-frontal.

4.4 Comparison of various classifiers

Until now, we only used linear 1-versus-1 SVMs for the classification. Here, we now compare different approaches. We selected a *nearest neighbor* (NN) classifier, a *nearest mean* (NM) classifier on top of *action unit* (AU) intensity estimates, and 1-versus-1 SVMs with linear, second-order polynomial, and RBF kernels. Here, we only tested the NN classifier for DCT features. The parameters for the SVM and the kernels were estimated using a 5-fold cross validation on folds which ensured that frames of the same subject end up in only one fold. For this parameter estimation, the slack parameter $C = 2^k$ was determined in the range $k = -10, \ldots, 0$. The $\gamma = 2^l$ for the RBF kernel was searched for in the range $l = -16, \ldots, -7$. Finally, for the polynomial kernel we used an offset of 1.0.

The AU intensity estimator used as input to the NM classifier is trained on the Bosphorus database [15, 16]. This model was then used to determine the AU intensities for the individual frames of the training and validation set of the EmotiW2013 dataset, which were then used as feature input for the NM classifier. In Figure 2, the means calculated per class over all frames of the sequences belonging to that class and normalized by the overall mean are visualized for the training set. Here, one can see that the mean already reflects somewhat the FACS definition of the prototypic expressions, e.g. Angry has maximum peaks for AU4 and AU24, Disgust has a maximum peak for AU9, Fear has maximum peaks for AU1 and AU4, Happy has maximum peaks for AU12 and AU6, Sad has a maximum peak for AU15, and Surprise has maximum peaks for AU26 and AU2 and the second highest peak for AU1.

The results for these different approaches on 64×80 faces and the official video-only baseline results using LBP-TOP and RBF SVM are presented in Table 6. RBF SVM seems to give an improvement for DCT over its linear variant, but not in case of Gabor. This might be due to the higher dimensional Gabor features (3200) and the comparatively low number of samples per class to generalize enough. Whereas for DCT the dimensionality is much lower (800) and therefore there are also enough samples to give a better generalization. Using a polynomial kernel improves even more for the DCT features compared to the RBF kernel, but for Gabor it only improves over a RBF but not a linear kernel. The NN classifier gives the worst performance. Surprisingly,

Table 3: Comparison of correct classification rates (in %) for eye-based alignment (eye) versus mixture-of-parts based alignment (MoP) and MoP alignment for only frames with valid eye detections (MoP-reduced) for DCT and Gabor features.

Features	Alignment	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Overall
DCT	eye	27.12	2.00	**20.37**	41.94	21.82	12.50	9.62	19.95
	MoP	**32.20**	10.00	7.41	40.32	18.18	12.50	**25.00**	21.21
	MoP (reduced)	28.81	10.00	12.96	50.00	10.91	6.25	17.31	19.95
Gabor	eye	30.51	**20.00**	14.81	**53.23**	23.64	10.94	15.38	24.49
	MoP	25.42	14.00	16.67	41.94	14.55	**17.19**	13.46	20.96
	MoP (reduced)	**32.20**	14.00	12.96	**53.23**	**27.27**	**17.19**	17.31	**25.51**

Table 4: Comparison of correct classification rates (in %) for various face representations using lin. SVMs

Features	Resolution	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Overall
DCT	64×80	27.12	2.00	20.37	41.94	21.82	12.50	9.62	19.95
	96×120	38.98	4.00	12.96	46.77	21.82	10.94	19.23	22.73
	128×160	30.51	4.00	14.81	46.77	16.36	10.94	7.69	19.44
LBP	64×80	30.51	10.00	18.52	46.77	**23.64**	14.06	21.15	23.99
	96×120	38.98	6.00	20.37	48.39	**23.64**	12.50	**23.08**	25.25
	128×160	28.81	12.00	**24.07**	41.94	16.36	10.94	19.23	22.22
Gabor	64×80	30.51	20.00	14.81	53.23	**23.64**	10.94	15.38	24.49
	96×120	**44.07**	**32.00**	9.26	48.39	16.36	**15.63**	15.38	**26.26**
	128×160	40.68	28.00	18.52	48.39	16.36	**15.63**	5.77	25.25
GDCT	64×80	37.29	8.00	12.96	**56.45**	18.18	9.38	19.23	23.74

Table 5: Correct classification rates (in %) for the frontal and non-frontal clips of the validation set

Pose	Features	Alignment	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Overall
frontal	DCT	eye	22.73	0.00	18.18	51.43	21.21	0.00	5.00	20.59
		MoP	18.18	5.26	9.09	45.71	27.27	15.79	20.00	22.94
		MoP-reduced	18.18	15.79	18.18	**57.14**	12.12	5.26	15.00	22.94
	LBP	eye	27.27	10.53	**27.27**	48.57	**33.33**	10.53	15.00	27.65
	Gabor	eye	**36.36**	**26.32**	18.18	51.43	30.30	10.53	20.00	**30.00**
		MoP	27.27	10.53	13.64	**57.14**	18.18	**21.05**	**25.00**	27.06
		MoP-reduced	22.73	10.53	18.18	**57.14**	**33.33**	15.79	15.00	28.24
non-frontal	DCT	eye	29.73	3.23	**21.88**	29.63	**22.73**	**17.78**	12.50	19.47
		MoP	**40.54**	12.90	6.25	33.33	4.55	11.11	**28.13**	19.91
		MoP-reduced	35.14	6.45	9.38	40.74	9.09	6.67	18.75	17.70
	LBP	eye	32.43	9.68	12.50	44.44	9.09	15.56	25.00	21.24
	Gabor	eye	27.03	**16.13**	12.50	**55.56**	13.64	11.11	12.50	20.35
		MoP	24.32	**16.13**	18.75	22.22	9.09	15.56	6.25	16.37
		MoP-reduced	37.84	**16.13**	9.38	48.15	18.18	**17.78**	18.75	**23.45**

Table 6: Correct classification rates (in %) for different classifiers

Features	Classifier	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Overall
LBP-TOP	RBF SVM	**44.00**	20.00	14.81	43.55	34.55	**20.31**	9.62	**27.27**
DCT	NN	28.81	32.00	18.52	16.13	12.73	3.13	**21.15**	18.43
	lin. SVM	27.12	2.00	**20.37**	41.94	21.82	12.50	9.62	19.95
	poly SVM	35.59	2.00	14.81	50.00	29.09	7.81	15.38	22.73
	RBF SVM	16.95	6.00	3.70	43.55	**49.09**	9.38	11.54	20.45
Gabor	lin. SVM	30.51	20.00	14.81	**53.23**	23.64	10.94	15.38	24.49
	poly SVM	30.51	20.00	14.81	46.77	16.36	7.81	19.23	22.47
	RBF SVM	35.59	12.00	7.41	46.77	25.45	7.81	11.54	21.46
AU Int.	NM	32.20	**38.00**	7.41	46.77	14.55	1.56	1.92	20.45

Table 7: Correct classification rates (in %) for different training sets evaluated on Val and Test set.

	Train set	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Overall
Val	Baseline	44.00	20.00	**14.81**	43.55	34.55	**20.31**	9.62	**27.27**
	EmotiW	**44.07**	**32.00**	9.26	48.39	16.36	15.63	**15.38**	26.26
	Google	40.68	18.00	0.00	**72.58**	**38.18**	3.13	11.54	27.02
Test	Baseline	**50.00**	**12.24**	0.00	48.00	18.75	6.97	5.71	22.75
	EmotiW	48.15	8.16	**12.12**	52.00	14.58	**13.95**	**20.00**	25.64
	Google	33.33	**12.24**	3.03	**86.00**	**43.75**	2.33	8.57	**29.81**

the very simple approach of using a NM classifier on top of the AU intensity estimates gives comparable results to using DCT with a RBF SVM. Thus, one could guess that with a more sophisticated machine learning approach than the NM classifier their could be even more gain in performance. Also, using Gabor instead of DCT features for the AU intensity estimation might add another boost. Compared to the video-only baseline using LBP-TOP and RBF SVM all the other approaches presented here are worse. This might be mainly due to the additional time information which is incorporated through the LBP-TOP features, although using Gabor features at a resolution of 96×120 comes close to the baseline without using any timing information.

4.5 Influence of Training Data

To see how a classifier performs when not trained on the EmotiW2013 dataset, but on some external data, we trained the classifiers alternatively on the web expression dataset provided by Richter et al. [14]. This dataset is also collected via a semi-automatic process using images from *Google Images*. The dataset consists of 4761 images labeled for the seven basic emotions also used in EmotiW2013.

The results for this experiment using linear 1-vs-1 SVMs and Gabor on 96×120 face images are shown in Table 7. We can see that the overall results using Google Images are even better than the once using the EmotiW training set when evaluating on the validation set and on the test set. This suggests that using all samples from a sequence as training data for a specific class deteriorates the performance, since the clips might contain also facial expressions not relevant for the labeled emotion. On the test set, both models even outperformed the baseline, which might suggest that either the timing information did not help as much on the test set or the baseline model was overfitted because of the use of a RBF kernel.

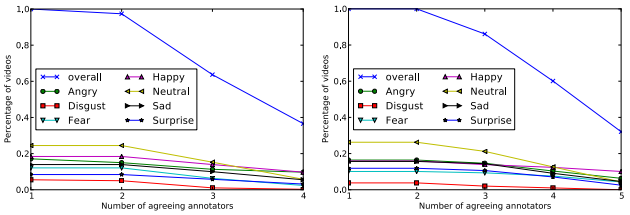
4.6 Human Evaluation

Since the baseline results on the EmotiW2013 dataset are very low, and our experiments also did not achieve better results on the validation set, we decided to do a human evaluation to compare to how well humans do in classifying the emotions of the clips and what the difficulties might be.

For this experiment, four, respectively five, persons labeled the training, respectively validation set. Since our approach is only using the video data we decided to let the annotators label the clips without listening to the audio track. Figure 3a and Figure 3b depict the percentage of clips from the training set, respectively validation set, the annotators agreed on with respect to how many annotators at least agreed. The figures show similar trends. Basically on almost all videos there was an agreement of at least two annotators. But three people (75%) agreed only on 63.7%

Table 8: Correct classification rates (in %) of human annotators on the video-only validation set.

Labeler	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Overall
human 1	71.19	10.00	48.15	69.35	70.91	53.13	65.38	56.31
human 2	67.80	20.00	40.74	82.26	**74.55**	48.44	21.15	52.02
human 3	62.71	22.00	55.56	**83.87**	65.45	57.81	55.77	58.59
human 4	**76.27**	6.00	46.30	79.03	56.36	**67.19**	**71.15**	58.84
human 5	62.71	**28.00**	**59.26**	82.26	63.64	62.50	50.00	**59.34**
Average	68.14	17.00	50.00	79.35	66.18	57.81	52.69	57.02



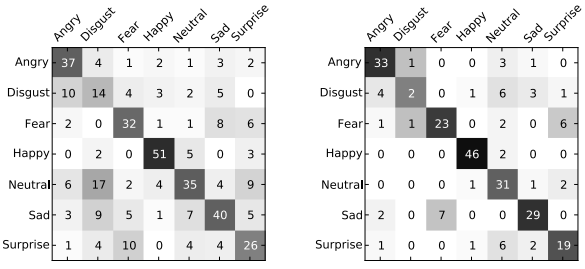
(a) training set (b) validation set

Figure 3: Percentage of clips for which at least a specific number of the human annotators agreed on when only taking the video stream into account.

of the training set and four people (80%) on 60.1% of the validation set. The inter-rater reliability in terms of Fleiss' (overall) kappa is 49.76% ($z = 55.26$, $p = 0$) for the training set and 52.63% ($z = 77.9$, $p = 0$) for the validation set and thus shows moderate agreement. Looking at the agreement for the individual classes, disgust is the least agreed on.

The results in terms of correct classification rate with respect to the official labels are presented in Table 8.

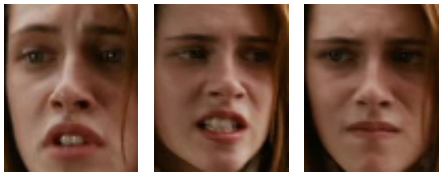
In Figure 4a, the corresponding confusion matrix for human 5 is shown. We see that even though humans are able to perform twice as good as the evaluated automatic approaches, the performance is still very bad. One thing that the human annotators noticed is that there are quite a few videos for which it is hard to decide the emotion class. Reasons for that are first of all, often it is hard because one needs more context knowledge to make a final decision, since the facial expression is too ambiguous or too subtle. As a human, one tries to do lip reading then or derive context knowledge from the interaction of the characters or events



(a) human 5 from Table 8 with respect to the official labels

(b) official labels with respect to the revised human annotated subset

Figure 4: Confusion matrices on the validation set. The rows correspond to the estimates, the columns to the reference.



(a) sad (b) anger (c) disgust

Figure 5: In clip 010025312 from the movie "Messengers", labeled with "sad" in the validation set, multiple emotions are present. The face crop-outs depict the different prototypic facial expressions. The character's spoken text is "You never listen to me!".

that happen in the background. But in some cases, this is not enough and one would rather additionally need the audio or even longer clips. In other cases, the labelers would have rather liked to assign some other class of emotion to the clip which is not one of the seven basic emotions, like contempt or frustration or the like. The lower agreement on disgust is due to the fact that it sometimes really seems to be more or less arbitrary if a clip is labeled as anger or disgust, or the expression is too subtle so that it gets confused with neutral. For other clips, it was hard to decide for which person the emotion had to be determined. This had two reasons, either there were shot changes within the clip so that over time different people were present in the clip, or in some cases there were even multiple persons visible at the same time. Sometimes it was even the case that one could see multiple different emotional expressions over time in one clip. This might be due to mixed feelings, like in the example in Figure 5. There, the character spoke the text "You never listen to me!" while looking throughout the clip at multiple people (not visible in the clip). First she shows sadness, but then moves to anger and finally to disgust. For this clip some of the annotators even labeled fear. When characters were talking, it was sometimes even more difficult to determine the emotion just from the facial expression, since then, the mouth movement, due to the speech, distorts the actual facial expression, but at other times the dynamics and gestures during the speech add additional context knowledge.

The reason for the much better performance of the humans compared to the automatic approaches is obviously the already mentioned context knowledge which humans can make use of and which is not incorporated in the approaches evaluated here. But also the different illumination conditions, wild pose changes, or occlusions which might lead to failures in the alignment or also in the classification.

4.7 Evaluation on revised subset

Since the official labels are only derived from one person, they might not have enough general agreement, as we saw that at least using video only, humans have a rather low agreement. Thus, we decided to derive a subset on which at least 75% of the human annotators agreed on. The results for some of the automatic approaches evaluated only on that revised subset are presented in Table 9, and the confusion matrix of the official labels with respect to this subset is given in Table 4b. It shows that there is still a big discrepancy between the revised human subset and the corresponding subset of the official labels for the disgust class. There is some slight increase in performance for the other emotions (besides neutral), but still the overall classification rate of 76.89% is a bit low for labels, but that might also be due to the fact that our human annotators had only the video track to annotate the data without context knowledge from the audio track or other sources. For the automatic approaches now the effect of the RBF kernel is reversed, i.e. for DCT the results drop while for Gabor they improve using a RBF kernel.

5. CONCLUSIONS

We presented results from extensive experiments, both in terms of machine and human performance. We showed that automatic facial expression analysis in the wild is still very challenging and that even humans only achieve an agreement of 52.63% in terms of Fleiss' kappa over all classes. The best result on the validation set using an automatic approach trained on the challenge training set, with 26.26% overall correct classification rate, was achieved for Gabor filtered images of 96×120 pixels and linear 1-versus-1 SVMs. Slightly better results, namely 27.02% overall correct classification rate, were achieved when training the same approach on an external dataset of web images. These results are relatively close to the baseline, which additionally makes use of temporal information via spatio-temporal features. When evaluated on the official test set, our best approach even outperformed the baseline by 7.06% absolute when using external data for training.

One of the problems we noticed is due to the nature of the labels of the dataset. Since there are only emotion labels for a whole clip available, but the clips might contain even multiple emotions or non-emotional faces, the classifiers trained on all frames might not work quite as well as if the frames would be properly filtered. To cope with such a situation multiple instance learning based approaches might work better [19, 18]. This also influences the estimation and evaluation, since to determine the emotion of a clip, one has to derive a label from all the frame estimates. When taking also non-emotional faces or even other emotions into account for the estimation of the clip estimate, the performance might again drop. On the other hand, since the clips are not presegmented to contain only one person with only one expression, sometimes sequences contain multiple different emotions either because the same person expresses them over time or because multiple persons in different emotional states are visible. To account for that in the evaluation procedure, the ground truth labels should actually also reflect that, and it should be either possible for the automatic approaches to choose one of these expressions contained in the clip or the evaluation should be rather performed on a frame basis. This is also related to the problem of segmenting the clip into parts with emotional content, which needs a completely different kind of benchmark.

6. ACKNOWLEDGMENTS

This work is funded by the "Concept for the Future" of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative.

We thank all the students and staff members who helped in the human evaluation: Cemal Çağrı Çörez, Daniel Koester, Timo Schneider, Makarand Tapaswi.

Table 9: Correct classification rates for different classifiers on revised human labels

Features	Classifier	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Overall
original labels		80.49%	50.00%	76.67%	93.88%	62.00%	80.56%	67.86%	76.89%
DCT	lin. SVM	31.71%	0.00%	**16.67%**	51.02%	34.00%	8.33%	3.57%	26.89%
DCT	RBF SVM	12.20%	0.00%	0.00%	48.98%	**52.00%**	5.56%	7.14%	24.79%
Gabor	lin. SVM	34.15%	**25.00%**	10.00%	**63.27%**	16.00%	16.67%	10.71%	27.73%
Gabor	RBF SVM	36.59%	0.00%	6.67%	59.18%	38.00%	**19.44%**	**14.29%**	**31.93%**
AU Int.	NM	**39.02%**	**25.00%**	3.33%	59.18%	12.00%	0.00%	0.00%	22.27%

7. REFERENCES

[1] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Automatic Recognition of Facial Actions in Spontaneous Expressions. *Journal of Multimedia*, 1(6):22–35, 2006.

[2] C.-C. Chang and C.-J. Lin. *LIBSVM: A Library for Support Vector Machines*, 2001. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

[3] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion Recognition In The Wild Challenge 2013. In *ACM International Conference on Multimodal Interaction*, 2013.

[4] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE Multimedia*, 19(3):34–41, 2012.

[5] H. K. Ekenel and R. Stiefelhagen. Analysis of Local Appearance-based Face Recognition: Effects of Feature Selection and Feature Normalization. In *Proc. of CVPR Biometrics Workshop*, 2006.

[6] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System - The Manual*. 2002.

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

[8] T. Gehrig and H. K. Ekenel. A common framework for real-time emotion recognition and facial action unit detection. In *CVPR Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB)*, 2011.

[9] T. Gehrig and H. K. Ekenel. Facial Action Unit Detection Using Kernel Partial Least Squares. In *1st IEEE Int. Workshop on Benchmarking Facial Image Analysis Technologies (BeFIT)*, 2011.

[10] C. Küblbeck and A. Ernst. Face detection and tracking in video sequences using the modified census transformation. *Image and Vision Computing*, 24(6):564–572, 2006.

[11] G. C. Littlewort, J. Whitehill, T.-F. Wu, N. Butko, P. Ruvolo, J. Movellan, and M. Bartlett. The motion in emotion — A CERT based approach to the FERA emotion challenge. In *FG Workshop on Facial Expression Recognition and Analysis Challenge (FERA)*, 2011.

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

[13] M. Pantic and L. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[14] M. Richter, T. Gehrig, and H. K. Ekenel. Facial Expression Classification on Web Images. In *Int. Conf. on Pattern Recognition*, 2012.

[15] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus Database for 3D Face Analysis. In *The 1st COST 2101 Workshop on Biometrics and Identity Management (BIOID)*, 2008.

[16] A. Savran, B. Sankur, and M. Taha Bilge. Regression-based Intensity Estimation of Facial Action Units. *Image and Vision Computing*, 30(10):774–784, 2012.

[17] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[18] K. Sikka, A. Dhall, and M. Bartlett. Weakly Supervised Pain Localization using Multiple Instance Learning. In *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2013.

[19] D. Tax, E. Hendriks, M. Valstar, and M. Pantic. The detection of concept frames using Clustering Multi-Instance Learning. In *Int. Conf. on Pattern Recognition*, 2010.

[20] Y.-I. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[21] M. F. Valstar, B. Jiang, M. Méhu, M. Pantic, and K. Scherer. The First Facial Expression Recognition and Analysis Challenge. In *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2011.

[22] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-Analysis of the First Facial Expression Recognition Challenge. *IEEE Trans. on Systems, Man, and Cybernetics, B: Cybernetics*, 42(4):966–979, 2012.

[23] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[24] X. Zhu and D. Ramanan. Face Detection, Pose Estimation, and Landmark Localization in the Wild. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.