Recognizing Expressions in a New Database Containing Played and Natural Expressions

James Skelley[†] Robert Fischer[†] Arup Sarma[†]
Bernd Heisele [‡†]
[†]Center for Biological and Computational Learning, M.I.T., Cambridge, USA
[‡]Honda Research Institute USA, Cambridge, USA
jskelley@alumn.mit.edu rfv@mit.edu arup@mit.edu bheisele@honda-ri.com

Abstract

We describe a new expression database which contains video sequences of both played and natural expressions and an expression classification system based on warped optical flow fields and texture features. We analyze the system's generalization performance when confronted with subjects that were not present in the training set and its recognition performance when tested on natural expressions. We evaluate several techniques for combining the classifier outputs computed on single images to perform classification of a temporal sequence of expression images.

1. Introduction

As the abilities of man-machine interfaces (MMI) to interact with humans approach a level on par with typical human-human interactions, the need for MMIs to recognize the same affective communication cues which humans use when conversing becomes increasingly apparent. In particular, it will be necessary for computers to recognize the same facial-emotional cues which humans use to direct the course of a conversation in the course of their interactions with the user.

Although much effort has been devoted to developing methods for automatic facial feature analysis and emotion recognition, only little attention has been paid to collecting appropriate training and test data. In [16], generating a comprehensive and labeled training database of expressions has been stated as one of the main challenges in automatic facial expression recognition. Because of the lack of appropriate databases, researchers in expression recognition frequently use databases developed for face recognition [17, 19] and person identification [21]. Unfortunately, these databases usually don't contain video sequences and also lack labeling information for expressions. A second set of widely used expression databases stems from behavioral science [11, 7]. These databases often contain specific facial muscle contractions but no natural expressions. In any of the expression databases which we evaluated, we discovered at least one of the following deficiencies: (a) the image quality was poor, (b) the number of expression samples per class was too small, (c) the database did not include video sequences, (d) the database did not include natural expressions, (e) the label information was insufficient. This prompted the construction of a new database which specifically addressed the issues listed above.

There are many challenges to automated facial expression recognition within an MMI setting. One problem is pose invariance. Slight changes in the pose of the face can be corrected by an alignment step prior to extracting the features for classification. Aligning a face to a canonical view requires the computation of correspondences between a reference image and the image to be aligned. The correspondences can be computed sparsely, i.e. by matching or tracking a few facial features (e.g. [12, 2]), or across large parts of the face (e.g. [25]). We applied an alignment algorithm which computes a dense correspondence map based on optical flow fields as suggested in [3].

An open problem is how to exploit temporal information for expression recognition. It can be addressed on various levels of complexity. For example, the results of the independently classified images can be merged to classify a whole image sequence. A more sophisticated approach is to model state transitions in expressions by hidden Markov models (e.g. [13, 6]). The comparison study between single frame and sequence analysis in [6] shows, somewhat surprisingly, no significant improvements when switching from single images to sequences. Our basic recognition module operates on single images. We extended it to incorporate temporal information by merging classification results across segmented expression sequences, i.e. sequences of only one expression.

Yet another problem is generalization: the system should be able to classify the expressions of a person who was not part of the training set. Expressions between people might vary due to differences in the shape and texture of their faces, and differences in the way they perform facial expressions. Texture dependencies can be removed by using optical flow as input to the classifier (e.g. [1, 15]). A technique to remove the shape dependency is multi-linear modeling to separate expression and identity as varying components within a tensor [22]. We try to avoid identity dependencies by warping the optical flow field of a persons expression onto a synthetic reference face image. This idea originates from work on morphable models, it has been applied to pose-invariant face identification [3, 23], visual speech synthesis [8], and expression mapping within a 3D morphable model framework [4].

Finally, the subtleness of natural expressions is a major challenge to recognition systems. Because of the lack of appropriate databases there are only very few studies on automatic recognition of natural expressions (e.g. [2, 18]).[1] In this paper we perform a preliminary investigation on how our system performs on natural expressions when trained on played expressions.

2. The Database

The facial expression database was intended for expression recognition systems within an MMI setting.[2] In this setting, the person will be close to the camera, either looking straight into the camera or at an object close by, e.g. a computer screen, or kiosk terminal. It is likely that head motion occurs in and out of the image plane when the user interacts with a computer. Thus, we did not constrain the head motion of the subjects during our recordings.

Two digital video cameras were used for recording frontal and half-profile views (about 30°) of the subjects. The scene was illuminated by a ceiling light and a photographic lamp with a diffuser, which was placed behind the camera at a height of approximately 2 meters.

The database consists of two parts: Played and natural expressions. In the first part, twelve subjects were asked to display six basic emotions: happiness, sadness, fear, disgust, surprise and anger. Data from these categories were each labeled as such. Because we intended to create a database demonstrating the widest gamut of expressions, we found this labeling more suitable than a more granular degree or valence labeling. Furthermore, degree-based labeling can be ambiguous, considering the range of values and unique responses provided by each of the subjects.

[1] To our knowledge, the databases for both experiments are not available to the public.

[2] A more detailed description is available in [20].

Most of the subjects had prior experience in acting. We asked the subjects to repeat each expression between 10 and 20 times in order to acquire a sufficiently large number of examples. In order to obtain natural expressions, the subjects watched a 20 minute movie that consisted of short video clips. The clips covered a large variety of topics, such as funny commercials, scenes of surgery and war documentaries. Similar to the HID database, the labeling of the recorded natural expressions was done by the judgment of the experimenter. Whenever possible, single expressions were isolated and cut from neutral state to neutral state. Sample images of the expression database are shown in figure 2. It should be noted that while the images of subject 103 are made available in the database, her data was ultimately not used during our recognition experiments since several of her image sequences contained expressions which were ambiguous and could not be properly labeled.

Figure 1. Sample images from our expression database. Images are provided for each subject (from left to right and top to bottom): 100, 101, 102, 103, 104, 105, 106, 107, 500, 501, 502, 503 respectively.

The labeled video clips were converted into sequences of images and then preprocessed as shown in figure 2. In step (1), each of the video clips were converted to sequences of images. In step (2), a face detector [24, 9] was used to localize the face in the image and to extract a region around the face in each image. To account for in-plane rotations of the subject's head, each of the images was rotated between $-30°$ and $+30°$ in steps of $5°$. We cropped the face at the location and rotation for which the face detector returned the greatest value and then rescaled the extracted face to a fixed size. In step (3), a neutral reference image was chosen for each of the twelve subjects and a binary mask was drawn on this image. A similarity transform was computed to approximate the mapping given by the optical flow within the masked region of the subject's reference image and a given

image from the subject's corpus of images. This transform was then applied to the corpus image to align it with the subject's reference image. The process was repeated for every image in the corpus. Thus, by the end of stage (4), each of the subjects has had the entirety of the images in his or her corpus aligned to a reference image particular to that subject. It was still necessary to align each of the subject's corpora to each other so that *all* images in the database were aligned to one another. In stage (5), this was accomplished by manually selecting four points in a synthetic reference face image.[3] By selecting the equivalent four points in the reference images of all the subjects, a second similarity transformation which would minimize the error between those corresponding points was computed. Each of the twelve transformations for each of the twelve subjects was then applied to the images of their respective corpora so that all images of the database were now aligned with the synthetic reference image. Before we extracted the features for classification, the images were masked as shown in step (6) to extract only the expression-relevant portions of the face. The same mask was applied to all images at exactly the same location without any manual interaction. Tables 1 and 2, delineate the final state of the database after preprocessing.

The described preprocessing procedure involved several manual steps. We could have chosen to fully automate this procedure, however, our main goal was to generate a clean database which would allow researchers to test classification algorithms without having to address the issues of accurate face detection and face alignment.

3. Features and Classifiers

A combination of two different feature types was employed.

The first set of features contained optical flow vectors [14] describing the changes in the shape of a person's face when performing an expression. To remove the components of the optical field containing subject-specific information, we performed a warping of the optical flow fields onto the synthetic reference face image which we used previously to align the faces.

We first computed the optical flow between this synthetic image and the reference image of each subject. Because both the synthetic image and the subject exhibit neutral expressions, the optical flow represented deformations related to the change in identity between the two faces. This flow is shown on the left of figure 3 and is referred to as optical flow (1). A second set of flow fields was computed between a subject's reference image and each of the images of this subject. The computation of this field is shown on the right of of figure 3 and referred to as optical flow (2). To re-

[3]The synthetic image was acquired from the website of the University of Regensburg in Germany [5]. The manually selected points were the center of each eye, the midpoint between the nostrils, and the center meeting point between the lips.

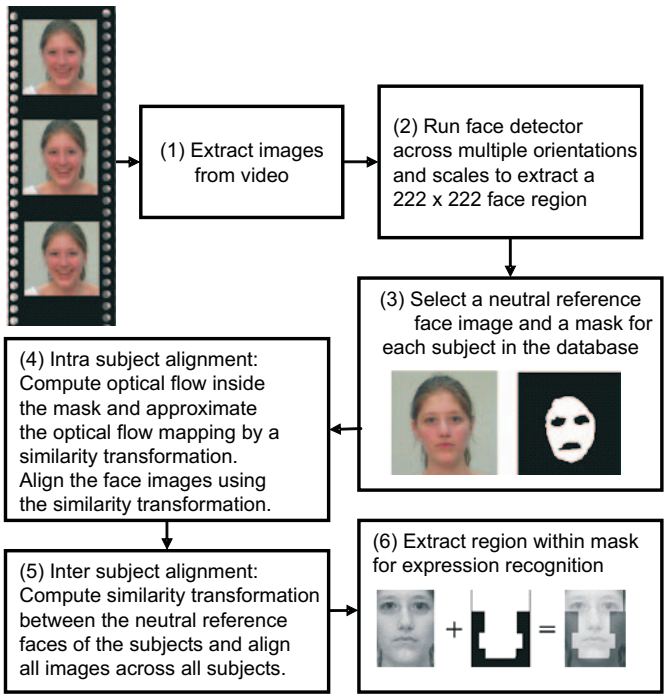

Figure 2. Preprocessing of the images.

subject	neutral	happiness	surprise	fear	sadness	disgust	anger
100	1200 / 6	609 / 9	324 / 11	929 / 16	1363 / 9	861 / 13	958 / 14
101	1800 / 10	1242 / 11	321 / 7	418 / 8	521 / 5	687 / 11	702 / 8
102	1260 / 7	1491 / 17	626 / 11	1180 / 14	1989 / 15	949 / 14	503 / 9
103	774 / 11	1301 / 14	1234 / 23	1713 / 25	1464 / 11	1693 / 15	1553 / 12
104	1080 / 6	963 / 14	507 / 19	927 / 18	1590 / 14	582 / 9	1204 / 19
105	617 / 9	1027 / 14	749 / 24	748 / 16	1822 / 10	483 / 9	1145 / 12
106	1260 / 7	1231 / 16	391 / 15	330 / 12	1904 / 16	328 / 9	627 / 18
107	1260 / 7	1046 / 24	1017 / 43	1006 / 33	717 / 18	931 / 27	436 / 11
500	900 / 5	1065 / 12	418 / 8	620 / 12	1253 / 14	874 / 11	1478 / 17
501	538 / 6	979 / 13	498 / 20	336 / 10	1045 / 12	400 / 8	442 / 8
502	200 / 2	775 / 12	658 / 14	556 / 14	647 / 7	456 / 9	612 / 13
503	1080 / 6	817 / 12	853 / 19	1238 / 10	895 / 11	953 / 13	1176 / 16

Table 1. Number of played expression images / sequences available per subject after preprocessing.

subject	smile	laugh	surprise	fear	shock	disgust	dislike	puzzlement
100	553 / 5	513 / 5	81 / 2	431 / 4	34 / 1	281 / 3	245 / 4	636 / 6
101	670 / 5	853 / 6	69 / 1	245 / 1	0 / 0	331 / 3	129 / 1	549 / 5
102	185 / 2	1408 / 11	0 / 0	0 / 0	0 / 0	105 / 2	386 / 3	81 / 1
103	200 / 3	856 / 5	340 / 6	0 / 0	0 / 0	153 / 3	520 / 5	70 / 1
104	72 / 1	2223 / 11	826 / 5	0 / 0	0 / 0	1222 / 6	0 / 0	86 / 1
105	88 / 1	980 / 11	87 / 1	0 / 0	135 / 2	440 / 6	333 / 5	123 / 2
106	202 / 2	870 / 6	203 / 2	171 / 1	125 / 1	281 / 2	535 / 4	540 / 5
107	308 / 3	360 / 4	0 / 0	0 / 0	270 / 4	126 / 2	249 / 3	49 / 1
500	288 / 2	1447 / 8	0 / 0	0 / 0	0 / 0	317 / 2	51 / 1	374 / 3
501	182 / 2	1026 / 10	115 / 2	0 / 0	0 / 0	95 / 2	126 / 2	0 / 0
502	287 / 3	790 / 6	0 / 0	0 / 0	0 / 0	296 / 3	574 / 6	100 / 1
503	308 / 2	1258 / 6	225 / 3	0 / 0	0 / 0	582 / 3	546 / 4	523 / 5

Table 2. Number of natural expression images / sequences available per subject after preprocessing.

move identity related information we mapped optical flow (2) onto the synthetic face. This mapping was achieved by using optical flow (1) to index into the values of optical flow (2), which can be interpreted as a warp of optical flow field (2) onto the synthetic reference image. The optical flows resulting from this warp operation were then masked as described in step (6) of figure 2. These masked flows were used as features by the expression classifiers.

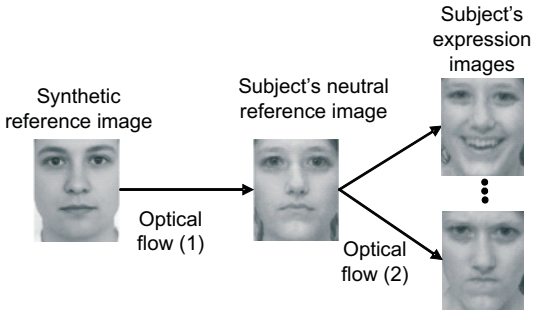

Figure 3. The optical flow (2) is warped onto the synthetic reference image using optical flow (1).

A second set of features describing the texture of a face was computed by applying the same mask used for the optical flow features to crop a part of the face from each image. Histogram equalization was then performed on the gray values of the extracted part to remove variations caused by lighting. The resulting gray values of the pixels were combined into a feature vector. Because the ranges of these two types of features did not agree, it was necessary to normalize them. This was accomplished by normalizing each feature set by its mean and variance.[4] After normalization, the two feature sets were merged into a single feature vector which was input to support vector machines (SVM) with linear kernels trained in a one-vs.-all (OVA) scheme. To classify sequences of expressions we combined the outputs of the single image classifiers belonging to the same expression sequence using the following strategies [10]:

- Maximum product:
  For each of the expression classes we computed the product of the real-valued classifier outputs across the sequence and then selected the class with the largest product. Prior to taking the product we normalized each classifier's output value by the softmax function.

- Maximum sum:
  For each of the expression classes we computed the sum over the real-valued classifier outputs and then selected the class with the largest sum. Prior to computing the sum we normalized each classifier's output values by softmax function.

- Majority vote:
  We computed the majority vote amongst the discrete-valued classifier outputs across the sequence. In the case of a tie we decided based on the maximum product.

[4]The particular set used for the calculation of the mean and variance was the training set for the group experiment described in section 4.2.

4. Experimental results

A completed expression recognition system, deployed in field, would ideally be able to:

- Recognize natural expressions having trained on played expression.

- Recognize expressions of individuals other than those on which it was trained.

Thus, a set of tests were run to verify the applicability of our system.[5] In the following tables, the label "Single" refers to SVMs trained on single images, "Sum" refers to sequence classification based on the maximum sum of the classifier outputs across the single images of a sequence, "Product" refers to sequence classification based on the maximum product of the classifier outputs across the single images, and finally "Majority" refers to sequence classification based on the majority vote across the single images.

4.1 Individual experiments: played vs. natural

Four expressions were common to both played expressions and natural expressions. These four expressions were trained upon using played data, and, as was available per subject, tested upon using the natural expressions. Recall that *all* available images for each natural expression were used for testing, in light of the paucity of images. The results are shown in table 3.[6]

The recognition rates are lower than in previous experiments where we achieved an average recognition rate of 92% on the same database with the same classifier when we trained *and* tested on played expressions [20]. This clearly shows that natural expressions are substantially more difficult to classify than played expressions. This is an important observation, considering that most recognition results in the literature have been reported on played expressions.

The experiments also show that combining the outputs of the single image classifier can boost recognition by over

[5]Experiments on training and testing person specific classifiers on played expressions only can be found in [20].

[6]When classifying sequences, we assigned the expression category determined for the full sequence to each image belonging to that sequence.

	Single	Sum	Product	Majority
Average rec. rate across subjects	50.5% / NA	65.7% / 62.2%	40.0% / 48.6%	64.7% / 60.8%
Standard dev. across subjects	18.3% / NA	12.5% / 16.9%	17.3% / 16.6%	17.5% / 19.2%

Table 3. Results for training and testing on the same subject's played and natural expressions for single image classification and different combination strategies for sequence classification. The results are given per image / per sequence.

Figure 4. Side-by-side demonstration of optical flow failure. Shown are three pairs of real and synthetic face images. The expression of the real face is warped onto the synthetic face image using optical flow. The optical flow does not properly capture the opening of the mouth in the rightmost pair of images.

10% using the majority voting and sum strategies. We conclude that the errors across a sequence must have been highly correlated, otherwise the increase in the recognition rate would have been significantly larger.

4.2 Group experiment: played vs. played

To evaluate how well the system can generalize when tested on subjects which were not in the training database, we performed the following experiments. The played expressions of five subjects (100, 101, 102, 502, 503) were trained upon, and then tested against the played expressions of six subjects (104, 105, 106, 107, 500, 501).

	Single	Sum	Product	Majority
Rec. rate	58.2% / NA	57.2% / 58.2%	48.4% / 54.9%	65.5 % / 60.7%

Table 4. Recognition rates for the group experiment for different combination strategies of image sequence classification. The results are given per image / per sequence.

Again, the recognition rates are considerably lower than in a previous experiment where we trained and tested on the same group of five people. The single image recognition rate in this experiment was 94% [20]. Despite our effort of removing the identity related components from the optical flow features the recognition rate drops by over 30% when tested on subjects which were not in the training set. There are two possible explanations for that: (a) the optical flow is faulty or the warping of the optical flow field onto the reference face does not completely remove identity related features. An example of a faulty optical flow is shown in figure 4: the optical flow does not correctly capture the opening of the mouth. And (b), the expressions vary significantly across different subjects and the training group of five subjects was too small to cover these variations.

The evaluations across sequences can improve the recognition rate by up to 7% when using the majority voting strategy. As in the previous experiment, the relatively small increase can be explained by a high correlation between the errors within a sequence.

5. Conclusion

We have described a new expression database, which includes high-quality video sequences of both played and natural expressions. Besides the original sequences, the database contains the sets of accurately extracted and aligned faces. This will allow researchers to skip the tedious detection and alignment stages and focus on the recognition task. Two types of features—optical flow and texture features—have been combined for expression recognition in a set of experiments using SVM-based classifiers. We investigated different techniques for combining the classification results computed on single images to perform classification of segmented expression sequences. From our trials we concluded that:

- Natural expressions are much more subtle than played expressions and will challenge a classifier's ability to recognize minute changes. Our experiments clearly show the necessity of databases of natural expressions with unconstrained head movements for evaluating the performance of real-world applicable expression recognition systems.

- Generalizing data from one set of individuals to another proved to be difficult despite the fact that we used warping of optical flow fields to remove identity related components.

- The best of the three investigated strategies for combining the single image classification results to perform expression sequence classification was the majority voting strategy. Improvements over single image classification were around 10%.

References

[1] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophysiology*, 36:253–263, 1999.

[2] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 568–573, 2005.

[3] D. Beymer. *Pose-invariant Face Recognition Using Real and Virtual Views*. PhD thesis, Massachusetts Institute of Technology, EECS, 1995.

[4] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In *Proc. of Eurographics*, volume 22, pages 641–650, 2003.

[5] C. Braun, M. Gruendl, C. Marberger, and C. Scherber. Beautycheck - Ursachen und Folgen von Attraktivitaet, 2001. University of Regensburg.

[6] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, 91(1-2):160–187, 2003.

[7] E. Douglas-Cowie, R. Cowie, and M. Schroeder. A new emotion database: considerations, sources and scope. In *Proc. of the ISCA Workshop on Speech and Emotion*, pages 39–44, 2000.

[8] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Proc. of ACM SIGGRAPH*, pages 388–398, 2002.

[9] The Intel Corporation. *OpenCV Online Reference Manual*, 2005.

[10] Y. Ivanov, B. Heisele, and T. Serre. Using component features for face recognition. In *Proc. of the 6th International Conference on Automatic Face and Gesture Recognition*, pages 421–426, 2004.

[11] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proc. of the 4th International Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000.

[12] A. Lanitis, C. J. Taylor, and T. F. Cootes. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):61:38–59, 1995.

[13] J. J. Lien. Automatic recognition of facial expressions using hidden markov models and estimation of expression intensity. Technical Report CMU-RI-TR-98-31, Carnegie Mellon University, 1998.

[14] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. DARPA Image Understanding Workshop*, pages 121–130, 1981.

[15] K. Mase. Recognition of facial expression from optical flow. *IEICE Transactions*, E74(10):3474–3483, 1991.

[16] J. Movellan and M. Bartlett. The next generation of automatic facial expression measurement. In P. Ekman, editor, *What the Face Reveals*. Oxford University Press, 2003.

[17] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi. The FERET evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[18] N. Sebe, M. S. Lew, I. Cohen, Y. Sun, T. Gevers, and T. S. Huang. Authentic facial expression analysis. In *Proc. of the 6th International Conference on Automatic Face and Gesture Recognition*, pages 517–522, 2004.

[19] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003.

[20] J. Skelley. Experiments in expression recognition. Master's thesis, Massachusetts Institute of Technology, EECS, 2005.

[21] A. J. Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, and H. Abdi. A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):812–816, 2005.

[22] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles. In *Proc. of the European Conference on Computer Vision (ECCV'02)*, pages 447–460, 2002.

[23] T. Vetter. Recognizing faces from a new viewpoint. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 139–144, 1997.

[24] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 511–518, 2001.

[25] L. Williams. Perfromance-driven facial animation. In *Proc. of ACM SIGGRAPH*, pages 235–242, 1990.