

Collecting Large, Richly Annotated Facial-Expression Databases from Movies

Abhinav Dhall

Australian National University

Roland Goecke

University of Canberra, Australia

Simon Lucey

Commonwealth Scientific and Industrial Research Organization (CSIRO)

Tom Gedeon

Australian National University

Two large facial-expression databases depicting challenging real-world conditions were constructed using a semi-automatic approach via a recommender system based on subtitles.

Collecting richly annotated, large datasets representing real-world conditions is a challenging task. With the progress in computer vision research, researchers have developed robust human facial-expression analysis solutions, but largely only for tightly controlled environments. Facial expressions are the visible facial changes in response to a person's internal affective state, intention, or social communication. Automatic facial-expression analysis has been an active research field for more than a decade, with applications in affective computing, intelligent environments, lie detection, psychiatry, emotion and paralinguistic

communication, and multimodal human-computer interface (HCI).

In the automatic human facial-analysis domain, realistic data plays an important role. However, as anyone in the facial-analysis community will attest, such datasets are extremely difficult to obtain. Much progress has been made in the facial- and human-activity-recognition fields in the past few years due to the availability of realistic databases as well as robust representation and classification techniques. However, although several popular facial-expression databases exist, the majority have been recorded in tightly controlled laboratory environments, where the subjects were asked to generate certain expressions. These lab scenarios are in no way a true representation of the real world. Ideally, we want a dataset of spontaneous facial expressions in challenging real-world environments.

To address this problem, we have collected two new facial-expression databases derived from movies via a semiautomatic recommender-based method. We extracted a database of temporal and static facial expressions from scenes in movies, environments that more closely resemble the real world than those of previous datasets. The database contains videos showing natural head poses and movements, close-to-real-world illumination, multiple subjects in the same frame, occlusions, and searchable metadata. The datasets also cover a large age range, including toddler, child, and teenager subjects, which are missing in other currently available temporal facial-expression databases.

Inspired by the Labeled Faces in the Wild (LFW) database,¹ we call our temporal database Acted Facial Expressions in the Wild (AFEW) and its static subset Static Facial Expressions in the Wild (SFEW).² In this context, "in the wild" refers to the challenging conditions in which the facial expressions occur rather than spontaneous facial expressions. We believe that these datasets will help advance facial-expression research and act as a benchmark for experimental validation of facial-expression analysis algorithms in real-world environments.

Constructing Facial-Expression Datasets

Until now, researchers have manually collected all facial-expression databases, which is time consuming and error prone. To address this limitation, we propose a video clip

recommender system based on subtitle parsing. Rather than manually scan a full movie, our labelers reviewed only the video clips suggested by the recommender system, which searched for clips with a high probability of a subject showing a meaningful expression. This method lets us collect and annotate large amounts of data quickly. Based on the availability of detailed information regarding the movies and their content on the Web, the labelers then annotated the video clips with dense information about the subjects. We used an XML-based representation for the database metadata, which makes it searchable and easily accessible using any conventional programming language.

Over the past decade, researchers have developed robust facial-expression analysis methods, which along with their different databases have followed various experimental protocols. This severely limits the ability to objectively evaluate the different methods. In response, we have defined clear experimental protocols, which represent different subject dependency scenarios.

Given the huge amount of video data on the Web, it is worthwhile to investigate the problem of facial-expression analysis in tough conditions. For the AFEW dataset, we labeled the video clips with one of six basic expressions: anger, disgust, fear, happiness, sadness, surprise, or neutral. The database captures facial expressions, natural head pose movements, occlusions, subjects' races, gender, diverse ages, and multiple subjects in a scene. Our baseline results show that current facial-expression recognition approaches that have reportedly achieved high recognition rates on existing datasets cannot cope with such realistic environments, underpinning the need for a new database and further research.

Although movies are often shot in somewhat controlled environments, they are significantly closer to real-world environments than current lab-recorded datasets. We do not claim that AFEW is a spontaneous facial-expression database. However, clearly, method actors attempt to mimic real-world human behavior to give audiences the illusion that they are behaving spontaneously, not posing, in movies. The AFEW dataset, in particular, addresses the issue of temporal facial expressions in difficult conditions that are approximating real-world conditions, which provides for a much

more difficult test set than currently available datasets.

Related Databases

One of the earliest databases published is the widely used Cohn-Kanade database,³ which contains 97 subjects who posed in a lab situation for the six universal and neutral expressions. Its extension CK+ contains 123 subjects, but the new videos were shot in a similar environment.³ The Multi-PIE database is another popular database that contains both temporal and static samples recorded in the lab over five sessions.⁴ It contains 337 subjects covering different pose and illumination scenes. Each of these databases were constructed manually, with the subjects posing in sequential scenes. The MMI database is a searchable temporal database with 75 subjects.⁶ All of these are posed, lab-controlled environment databases. The subjects display various acted (not spontaneous) expressions. The recording environment is nowhere near real-world conditions.

The RU-FACS (Rutgers and University of California, San Diego, Facial Action Coding System [FACS]) database is a FACS-coded temporal database containing spontaneous facial expressions,⁶ but it is proprietary and unavailable to other researchers. The Belfast database consists of a combination of studio recordings and TV program grabs labeled with particular expressions.⁷ The number of TV clips in this database is sparse. Compared to the manual method used to construct and annotate these databases, our recommender system method is faster and more easily accessible. The metadata schema is in XML and, hence, easily searchable and accessible from a variety of languages and platforms. In contrast, CK, CK+, Multi-PIE, RU-FACS, and Belfast must be searched manually.

The Japanese Female Facial Expression (JAFPE) database is one of the earliest static facial-expression datasets.⁸ It contains 219 images of 10 Japanese females. However, it has a limited number of samples and subjects and was also created in a lab-controlled environment.

In one of the first experiments on close-to-real data, Marco Paleari, Ryad Chellali, and Benoit Huet proposed a bimodal, audio-video features-based system.⁹ The database was constructed from TV programs, but it is fairly small, with only 107 clips.

Table 1. Comparison of temporal facial expression databases.*

| Database | Construction | | Age range | Illumination | Occlusion | Subjects | Searchable | Subject details | Multiple subjects |
|-------------------------|--------------|-------------|-----------|--------------|-----------|----------|------------|-----------------|-------------------|
| | process | Environment | | | | | | | |
| AFEW | Assisted | CTR | 1–70 | CTN | Yes | 330 | Yes | Yes | Yes |
| Belfast ⁷ | Manual | TV & Lab | ? | C | Yes | 100 | No | No | No |
| CK ³ | Manual | Lab | 18–50 | C | No | 97 | No | No | No |
| CK+ ³ | Manual | Lab | 18–50 | C | No | 123 | No | No | No |
| F.TUM ⁹ | Manual | Lab | ? | C | No | 18 | No | No | No |
| GEMEP ¹⁰ | Manual | Lab | ? | C | Yes | 10 | No | No | No |
| M-PIE ⁴ | Manual | Lab | 27.9 | C | Yes | 337 | No | No | No |
| MMI ⁵ | Manual | Lab | 19–62 | C | Yes | 29 | Yes | No | No |
| Paleari ¹¹ | Manual | CTR | – | CTN | Yes | – | No | No | No |
| RU-FACS ⁶ | Manual | Lab | 18–30 | C | Yes | 100 | No | No | No |
| Semaine ¹² | Manual | Lab | ? | C | Yes | 75 | Yes | No | No |
| UT-Dallas ¹³ | Manual | Lab | 18–25 | C | Yes | 284 | No | No | No |
| VAM ¹⁴ | Manual | CTR | ? | C | Yes | 20 | No | No | No |

* C stands for controlled, CTN for close to natural, and CTR for close to real.

Table 1 compares these and other facial-expression databases.

Our AFEW database is similar in spirit to the LFW database¹ and the Hollywood Human Actions (HOHA) dataset.¹⁵ These contain varied pose, illumination, age, gender, and occlusion. However, LFW is a static facial-recognition database created from single face images found on the Web specifically for face recognition, and HOHA is an action-recognition database created from movies.

Database Contributions

The AFEW and SFEW databases offer several novel contributions to the state of the art. AFEW is a dynamic, temporal facial-expression data corpus consisting of short video clips of facial expressions in close-to-real-world environments. To the best of our knowledge, SFEW is also the only static, tough conditions database covering the seven facial-expression classes.

Our subjects ranged from 1 to 70 years old, which makes the resulting datasets generic in terms of age, unlike other facial-expression databases. The databases have many clips depicting children and teenagers, which can be used to study facial expressions in younger subjects. The datasets can also be used for both static and temporal facial age research.

To the best of our knowledge, AFEW is currently the only facial-expression database with multiple labeled subjects in the same frame. This will enable interesting studies on various

themes (expressions) involving scenes with multiple subjects, who might or might not have the same expression at a given time.

The databases also exhibit close-to-real illumination conditions. The clips include scenes with indoor, nighttime, and outdoor natural illumination. Although movie studios use controlled illumination conditions, even in outdoor settings, these are closer to natural conditions than lab-controlled environments and, therefore, are valuable for facial-expression research. The diverse nature of the illumination conditions in the dataset makes it useful for not just facial-expression analysis but potentially also for facial recognition, facial alignment, age analysis, and action recognition.

The movies we chose cover a large set of actors. Many actors appear in multiple movies in the dataset, which will enable researchers to study how their expressions have evolved over time, whether they differ for different genres, and so forth.

The design of the database schema is based on XML. This enables further information about the data and its subjects to be added easily at any stage without changing the video clips. This means that detailed annotations with attributes about the subjects and the scene are possible.

The database download website will also contain information regarding the experiment protocols and training and test splits for both temporal and static facial-expression recognition (FER) experiments.

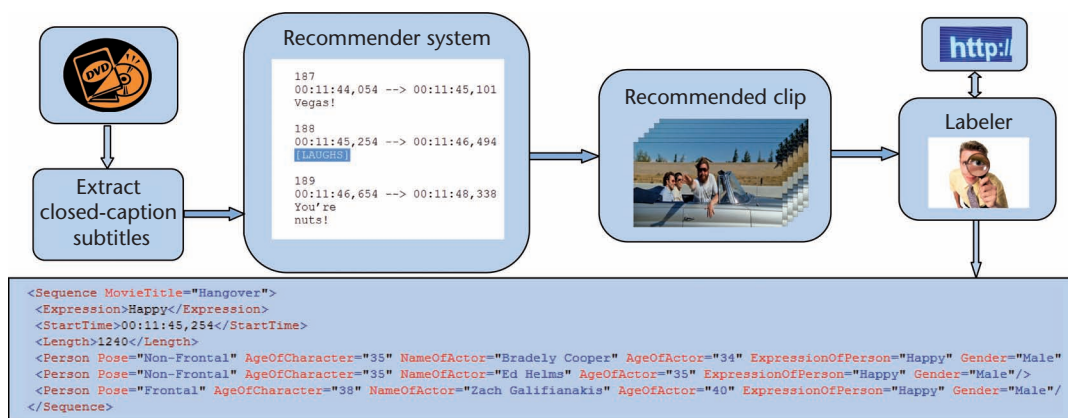


Figure 1. Database creation process. A subtitle is extracted from a DVD and then parsed by the recommender system. In this example, from the 2009 movie *The Hangover*, when the subtitle contains the keyword “laugh,” the tool plays the corresponding clip. The human labeler then annotates the subjects in the scene, using a GUI tool, based on the information about the subjects in the clip available on the Web. The resulting annotation, which in this case contains the information about a scene containing multiple subjects, is stored in the XML schema shown at the bottom of the diagram.

Database Creation

To construct the database, we followed a semi-automatic approach and divided the process into two parts (see Figure 1). First, the subtitles are extracted and parsed in the recommender system. Second, a human labeler annotates the recommended clips based on information available on the Internet.

Subtitle Extraction

We purchased and analyzed 54 movie DVDs. We extracted subtitles for the deaf and hearing impaired (SDH) and closed caption (CC) subtitles from the DVDs because they contain information about the audio and nonaudio context such as emotions and information about the actors and scene (for example, [CHEERING], [SHOUTS], and [SURPRISED]). We extracted the subtitles from the movies using the Vob-Sub Rip (VSRip) tool (www.videohelp.com/tools/VSRip). For the movies that VSRip could not extract subtitles, we downloaded the SDH from the Web. The extracted subtitle images were parsed using optical character recognition (OCR) and converted into the .srt subtitle format using the Subtitle Edit tool (www.nikse.dk/se). The .srt format contains the start time, end time, and textual content with millisecond accuracy.

Video Recommender System

Once the subtitles have been extracted, we parse the subtitles and search for expression-related keywords—for example, happy, sad,

surprised, shouts, cries, groans, cheers, laughs, sobs, silence, angry, weeping, sorrow, disappointment, and amazed. If found, the system recommends video clips to the labeler. The clip’s start and end time is extracted from the subtitle information. The system plays the video clips sequentially, and the labeler enters information about the clip and its characters and actors from the Web. If clips contain multiple actors, the labeling sequence is based on two criteria. For actors appearing in the same frame, the order of annotation is left to right. If the actors appear at different timestamps, then it is in the order of appearance. The dominating expression in the video is labeled as the *theme expression*. The labeling is then stored in an XML metadata schema. Finally, the labeler enters the character’s age or his or her estimated age if this information is unavailable.

In total, the subtitles from the 54 DVDs contained 77,666 individual subtitles. Out of these, the recommender system suggested 10,327 clips corresponding to subtitles containing expressive keywords. The labelers chose 1,426 clips from these on the basis of criteria such as the visible presence of subjects, at least some part of the face being visible, and the display of meaningful expressions.

Because subtitles are manually created by humans, they can contain errors. This might lead to a situation where the recommender system suggests an erroneous clip. However, the labelers can reject a recommendation. When

Table 2. AFEW database attributes.

| Attribute | Description |
|--|--|
| Length of sequences | 300–5,400 ms |
| Number of sequences | 1,426 |
| Total number of expressions (including multiple subjects) | 1,747 |
| Video format | AVI |
| Maximum number of clips of a subject | 134 |
| Minimum number of clips of a subject | 1 |
| Number of labelers | 2 |
| Number of subjects | 330 |
| Number of clips per expression | Anger (194), disgust (123), fear (156), sadness (165) happiness (387), neutral (257), surprise (144) |

annotating the clips, the labelers use the clips' video, audio, and subtitle information to make informed decisions. We can use the proposed recommender system to easily add more clips to the database and scale it up in the future.

Database Annotations

Our database contains metadata about the video clips in an XML-based schema, which enables efficient data handling and updating. The human labelers densely annotated the video clips with the expression and subject information.

The subject information contains various attributes describing the actor and/or character in the scene:

- *Pose*. This denotes the head pose based on the labeler's observation. In the current version, we manually classify the head pose as frontal or nonfrontal.
- *Character age*. Frequently, only the age of the lead actors' characters are available on the Web. The labeler estimated any other ages.
- *Actor name*. Here we provide the actor's real name.
- *Actor age*. The labelers extracted the actor's real ages from www.imdb.com. In a few cases, the age information was missing, so the labeler estimated it.
- *Expression of person*. This denotes the expression class of the character as labeled by the human observer. This could differ from the

higher-level expression tag because there might be multiple people in the frame showing different expressions with respect to each other and the scene/theme.

- *Gender*. Here we provide the actor's gender.

Expression tag specifies the theme expression conveyed by the scene. The expressions were divided into the six expression classes, plus neutral. The default value is based on the search keyword found in the subtitle text—for example, we use happiness for “smile” and “cheer.” Human observer can change it based on their observation of the audio and scene in the clip.

This XML-based metadata schema has two major advantages. First, it is easy to use and search using any standard programming language on any platform that supports XML. Second, the structure makes it simple to add new attributes about the video clips in the future, such as the pose of the person in degrees and scene information, while keeping the existing data and ensuring that pre-existing tools can exploit this information with minimal changes.

Currently, the database metadata indexes 1,426 video clips. Table 2 gives the database details. Additional information on how to obtain the database and its experimental protocols are available at <http://cs.anu.edu.au/few>.

SFEW

Static facial-expression analysis databases such as Multi-PIE and JAFFE are lab-recorded databases in tightly controlled environments. We extracted frames from AFEW to create a static image database that more closely represents the real world. Later, we describe the three versions of SFEW, which are based on the level of subject dependency for evaluating facial-expression recognition performance of systems in different scenarios. The strictly person-independent version of SFEW is described in an earlier work⁴ and is posted as a challenge on the Benchmarking Facial Image Analysis Technologies (BEFIT) website (<http://fipa.cs.kit.edu/511.php>).

Comparison with Other Databases

To evaluate our datasets, we compared the performance of state-of-the-art descriptors on AFEW and SFEW with that on existing, widely used datasets. Specifically, we compared AFEW to the CK+ database, which is an extension of the Cohn-Kanade database.³ A basic facial

expression consists of various temporal dynamic stages: onset, apex, and offset stage. In CK+, all videos follow the temporal dynamic sequence: *neutral* → *onset* → *apex*, which is not a true reflection of how expressions are displayed in real-world situations because the data about the offset phase is missing.

We also argue that all data containing the complete temporal sequence might not always be available. For example, a person entering a scene might already be happy and close to the highest intensity of happiness (onset). Earlier systems trained on existing databases like CK+ have learned on such stages. However, the availability of full temporal dynamic stages is not guaranteed in real-world settings. In our database, this is not fixed due to its close-to-natural settings. To extract a face, we computed the Viola-Jones detector¹⁶ over the CK+ sequences. In our comparison experiments, we used six common classes from both the AFEW and CK+ databases (anger, fear, disgust, happiness, sadness, and surprise).

We compared SFEW with the JAFFE and Multi-PIE databases in two experiments:

- a comparison of SFEW, JAFFE, and Multi-PIE on the basis of four common expression classes (disgust, neutral, happiness, and surprise) and
- a comparison of SFEW and JAFFE on all seven expression classes.

We computed feature descriptors on the cropped faces from all the databases. The cropped faces were divided into 4×4 blocks for local binary pattern (LBP),¹⁷ local phase quantization (LPQ),¹⁷ and pyramid histogram of gradients (PHOG).¹⁸ For LBP and LPQ, we set the neighborhood size to eight. For PHOG, bin length was eight, pyramid levels L were 2, and angle range equaled $[0, 360]$. We applied principal component analysis (PCA) on the extracted features and kept 98 percent of the variance. For classification, we used a support vector machine (SVM) learned model. The kernel was C-support vector classification (C-SVC), with a radial basis function (RBF) kernel. We used five-fold cross validation to select the parameters. For AFEW, the static descriptors were concatenated.

LBP-TOP performed the best out of all the methods. The overall expression classification accuracy is much higher for CK+ (see Figure 2).

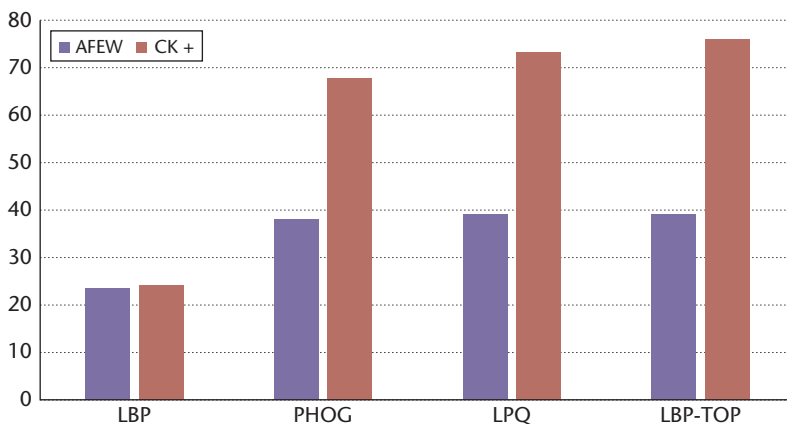


Figure 2. Performance of LBP, PHOG, LPQ, and LBP-TOP on the CK+ and AFEW databases. The descriptors performed poorly on the AFEW dataset.

For the SFEW four-expression class experiment, the classification accuracy on the Multi-PIE subset was 86.25 and 88.25 percent for LPQ and PHOG, respectively. For JAFFE, it was 83.33 percent for LPQ and 90.83 percent for PHOG. For SFEW, it was 53.07 percent for LPQ and 57.18 percent for PHOG. For the seven-expression class experiment, the classification accuracy for JAFFE was 69.01 percent for LPQ and 86.38 percent for PHOG. For SFEW, it was 43.71 percent for LPQ and 46.28 percent for PHOG. Thus, LPQ and PHOG achieve a high accuracy on JAFFE and Multi-PIE, but a significantly lower accuracy for SFEW.

In our opinion, the primary reason for the poor performance of state-of-the-art descriptors on AFEW and SFEW is that the databases on which these state-of-the-art methods have been experimented on were recorded in lab-based environments. Expression analysis in close-to-real-world situations is a nontrivial task and requires more sophisticated methods at all stages of the approach, such as robust face localization and tracking, illumination, and pose invariance.

Experimentation Protocols

Over the years, researchers have proposed many facial-expression recognition methods based on experiments on various databases following different protocols, making it difficult to compare the results fairly. Therefore, we created strict experimentation protocols for both databases. The different protocols are based on the level of person dependency present in the sets (see Table 3).

Table 3. Experimentation protocol scenarios for SFEW and AFEW.

| Protocol | AFEW/SFEW training-test content |
|-----------------------------------|--|
| Strictly Person Specific (SPS) | Same single subject |
| Partial Person Independent (PPI) | A mix of common and different subjects |
| Strictly Person Independent (SPI) | Different subjects ² |

The BEFIT workshop challenge² falls under Strictly Person Independent (SPI) for SFEW. Data, labels, and other protocols will be made available on the database website. AFEW Partial Person Independent (PPI) contains 745 videos and AFEW SPI contains 741 videos in two sets. AFEW Strictly Person Independent (SPI) contains 40 videos of the actor Daniel Radcliffe for four expression categories (fear, happiness, neutral, and surprise). For SFEW, SFEW SPS contains 76 images of Daniel Radcliffe for five expression classes (anger, fear, happiness, neutral, and surprise). SFEW PPI contains 700 images and SFEW SPI contains 700 images in two sets.

Baseline

For all the protocols for SFEW, we computed the baselines based on the method defined in our earlier work.² (These results are an average of training and testing on the sets.) PHOG and LPQ features were computed on the cropped face. The features were concatenated together to form a feature vector. For dimensionality reduction, we computed PCA and kept 98 percent of the variance. Furthermore, we used a nonlinear SVM to learn and classify expressions. (Again, see our earlier work for the parameter selection details.²) To encode the temporal data, we computed LBP-TOP features, as in the previous section.

Table 4 shows the classification accuracy for both databases and their protocols. The low

classification accuracy results demonstrate that the current methods are inappropriate for real-world scenarios.

Conclusions

Facial-expression analysis is a well-researched field. However, progress in the field has been hampered due to the unavailability of databases depicting real-world conditions. State-of-art FER methods that have performed well on existing datasets do not work well on the datasets we propose here. This is due to a lack of robust “in the wild” face-alignment methods and efficient temporal descriptors.

As part of future work, we will adapt current methods and extend algorithms for FER in tough conditions. AFEW contains group-level-expression video clips, which in the future can be used to develop systems that analyze theme expressions in scenes containing groups of people. We believe that these datasets will enable novel contributions to facial-expression research and act as a benchmark for experimental validation of facial-expression analysis algorithms in real-world environments. **MM**

References

1. G.B. Huang et al., *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*, tech. report 07-49, Univ. of Massachusetts, Amherst, 2007.
2. A. Dhall et al., “Static Facial Expression Analysis in Tough Conditions: Data, Evaluation Protocol and Benchmark,” *Proc. IEEE Int’l Conf. Computer Vision Workshop (BeFIT)*, IEEE Press, 2011, pp. 2106–2112.
3. P. Lucey et al., “The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPR4HB)*, IEEE CS Press, 2010, pp. 94–101.

Table 4. Average classification accuracies of different protocols.

| Protocol | Anger (%) | Disgust (%) | Fear (%) | Happiness (%) | Neutral (%) | Sadness (%) | Surprise (%) | Average (%) |
|----------|-----------|-------------|----------|---------------|-------------|-------------|--------------|-------------|
| AFEW PPI | 32.5 | 12.3 | 14.1 | 44.2 | 33.8 | 25.2 | 21.8 | 26.3 |
| AFEW SPI | 40.1 | 7.9 | 14.5 | 37.0 | 40.1 | 23.5 | 8.9 | 24.5 |
| AFEW SPS | – | – | 0.0 | 50.0 | 0.0 | – | 50.0 | 25.0 |
| SFEW PPI | 29.5 | 43.5 | 48.5 | 35.5 | 33.0 | 12.0 | 35.0 | 33.8 |
| SFEW SPI | 23.0 | 13.0 | 13.9 | 29.0 | 23.0 | 17.0 | 13.5 | 18.9 |
| SFEW SPS | 35.0 | – | 45.8 | 0.0 | 7.1 | – | 0.0 | 17.5 |


4. R. Gross et al., "Multi-PIE," *Proc. 8th IEEE Int'l Conf. Automatic Face and Gesture Recognition (FG)*, 2008, pp. 1–8, doi:10.1109/AFGR.2008.4813399.
5. M. Pantic et al., "Web-Based Database for Facial Expression Analysis," *Proc. IEEE Int'l Conf. Multimedia and Expo (ICME)*, IEEE CS Press, 2005, pp. 317–321.
6. M.S. Bartlett et al., "Automatic Recognition of Facial Actions in Spontaneous Expressions," *J. Multimedia*, vol. 1, no. 6, 2006, pp. 22–35.
7. E. Douglas-Cowie, R. Cowie, and M. Schröder, "A New Emotion Database: Considerations, Sources and Scope," *Proc. ISCA ITRW on Speech and Emotion*, 2000, pp. 39–44.
8. M.J. Lyons et al., "Coding Facial Expressions with Gabor Wavelets," *Proc. IEEE Int'l Conf. Automatic Face Gesture Recognition and Workshops (FG)*, IEEE CS Press, 1998, p. 200.
9. F. Wallhoff, "Facial Expressions and Emotion Database," 2006, www.mmk.ei.tum.de/~waf/fgnet/feedtum.html.
10. T. Bänziger and K. Scherer, "Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) Corpus," *Blueprint for Affective Computing: A Sourcebook*, K. Scherer, T. Bänziger, and E. Roesch, eds., Oxford Univ. Press, 2010.
11. M. Pleari, R. Chellali, and B. Huet, "Bimodal Emotion Recognition," *Proc. 2nd Int'l Conf. Social Robotics (ICSR)*, Springer, 2010, pp. 305–314.
12. G. McKeown et al., "The SEMAINE Corpus of Emotionally Coloured Character Interactions," *Proc. IEEE Int'l Conf. Multimedia and Expo (ICME)*, IEEE CS Press, 2010, pp. 1079–1084.
13. A.J. O'Toole et al., "A Video Database of Moving Faces and People," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, 2005, pp. 812–816.
14. M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database," *Proc. IEEE Int'l Conf. Multimedia and Expo (ICME)*, IEEE CS Press, 2008, pp. 865–868.
15. I. Laptev et al., "Learning Realistic Human Actions from Movies," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, IEEE CS Press, 2008, pp. 1–8.
16. P.A. Viola and M.J. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, IEEE CS Press, 2001, pp. 511–518.
17. D. Huang et al., "Local Binary Patterns and its Application to Facial Image Analysis: A Survey," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 41, no. 6, 2011, pp. 1–17.
18. A. Bosch, A. Zisserman, and X. Munoz, "Representing Shape with a Spatial Pyramid Kernel," *Proc. 6th ACM Int'l Conf. Image and Video Retrieval (CIVR)*, ACM Press, 2007, pp. 401–408.

Abhinav Dhall is a doctoral candidate in the Research School of Computer Science at the Australian National University. His research interests include affective computing, computer vision, pattern recognition, and human-computer interaction. Dhall has a BS in computer science from the DAV Institute of Engineering and Technology, India. He is a student member of IEEE and was awarded the 2010 Australian Leadership Award Scholarship. Contact him at abhinav.dhall@anu.edu.au.

Roland Goecke leads the Vision and Sensing Group in the Faculty of Information Sciences and Engineering at the University of Canberra, Australia. His research interests include affective computing, computer vision, human-computer interaction, and multimodal signal processing. Goecke has a PhD in computer science from the Australian National University. He is a member of IEEE. Contact him at roland.goecke@ieee.org.

Simon Lucey is a Senior Research Scientist in the CSIRO and a "Futures Fellow Award" recipient from the Australian Research Council. He holds adjunct Professorial positions at the University of Queensland and, Queensland University of Technology. His research interests include computer vision and machine learning and their application to human behavior. He is a member of IEEE and was awarded the 2009 ARC Future Fellowship by the Australian Research Council. Lucey has a PhD in computer science from the Queensland University of Technology, Brisbane. Contact him at simon.lucey@csiro.au.

Tom Gedeon is the chair professor of computer science at the Australian National University and president of the Computing Research and Education Association of Australasia. His research interests include the development of automated systems for information extraction and synthesis into humanly useful information resources, mostly using fuzzy systems and neural networks. Gedeon has a PhD from the University of Western Australia. He is a member of IEEE and a former president of the Asia-Pacific Neural Network Assembly. Contact him at tom@cs.anu.edu.au.

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.