A Semi-Automatic Method for Collecting Richly Labelled Large Facial Expression Databases from Movies

Abhinav Dhall, *Student Member, IEEE,* Roland Goecke, *Member, IEEE,*
Simon Lucey, *Member, IEEE,* and Tom Gedeon, *Member, IEEE*

Abstract

Creating large, richly annotated databases depicting real-world or simulated real-world conditions is a challenging task. There has been a long understood need for recognition of human facial expressions in realistic video scenarios. Although many expression databases are available, research has been restrained by their limited scope due to their 'lab controlled' recording environment. This paper proposes a new temporal facial expression database *Acted Facial Expressions in the Wild (AFEW)* and its static subset *Static Facial Expressions in the Wild (SFEW)*, extracted from movies. As creating databases is time consuming and complex, a novel semi-automatic approach via a recommender system based on subtitles is proposed. Further, experimental protocols based on varying levels of person dependency are defined. AFEW is compared with the extended Cohn-Kanade CK+ database and SFEW with JAFFE and Multi-PIE databases

Index Terms

Facial expression recognition, large scale database, real-world conditions, emotion database

A. Dhall, and T. Gedeon are with the College of Engineering and Computer Science, Research School of Computer Science, Australian National University, Canberra, Australia. Email: abhinav.dhall@anu.edu.au, tom.gedeon@anu.edu.au

R. Goecke is with the Faculty of Information Sciences and Engineering, University of Canberra, Canberra, Australia. Email: roland.goecke@ieee.org

S. Lucey is with the Commonwealth Scientific & Industrial Research Organisation, Brisbane, Australia. Email: simon.lucey@csiro.au

I. INTRODUCTION

With the progress in computer vision research, robust human facial expression analysis solutions have been developed 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 field of research for over a decade now. It finds applications in affective computing, intelligent environments, lie detection, psychiatry, emotion and paralinguistic communication, and multimodal human computer interface (HCI).

Image analysis is inherently data-driven. In the domain of automatic human face analysis, realistic data plays a very important role. Much progress has been made in the fields of face recognition and human activity recognition in the past years due to the availability of realistic databases as well as robust representation and classification techniques. However, in the case of human facial expression analysis, there is a lack of databases representing real-world scenarios. Even though there are a number of popular facial expression databases, the majority of these have been recorded in tightly controlled laboratory environments, where the subjects were asked to 'act' certain expressions. These 'lab scenarios' are in no way a (close to) true representation of the real world. Ideally, we would like a dataset of spontaneous facial expressions in challenging real-world environments. However, as anyone in the face analysis community will attest to, such datasets are extremely difficult to obtain.

As an important stepping stone on this path, we have collected a temporal and a static facial expression database extracted from scenes in movies (in environments significantly more closely resembling the real world than those of current widely used datasets). Inspired by the LFW database [9], we call the temporal database *Acted Facial Expressions in the Wild (AFEW)* and its static subset *Static Facial Expressions in the Wild (SFEW)* [4]. Note, 'in the wild' here refers to the challenging conditions in which the facial expressions occur, not to these being databases of spontaneous facial expressions recorded 'in the wild'. For the transition of facial expression analysis approaches from lab to realistic environments, we need databases that can mimic the real world.

Until now, all facial expression databases have been collected manually, which makes the process time consuming and error prone. To address this limitation, we propose a video clip recommender system, which is based on subtitle parsing. The labelers in our case did not scan the full movie manually but used the recommender system, which suggests only those video clips, which have 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 contents

Database	Const. Process	Environ.	Age Range	Illum.	Occl.	Subjects	Searchable	Subject Details	Multi. Subj.
AFEW	Assisted	CTR	1-70	CTN	Yes	330	Yes	Yes	Yes
Belfast [5]	Manual	TV & Lab	?	C	Yes	100	No	No	No
CK [11]	Manual	Lab	18-50	C	No	97	No	No	No
CK+ [11]	Manual	Lab	18-50	C	No	123	No	No	No
F.TUM [18]	Manual	Lab	?	C	No	18	No	No	No
GEMEP [1]	Manual	Lab	?	C	Yes	10	No	No	No
M-PIE [7]	Manual	Lab	27.9	C	Yes	337	No	No	No
MMI [16]	Manual	Lab	19-62	C	Yes	29	Yes	No	No
Paleari [15]	Manual	CTR	-	CTN	Yes	-	No	No	No
RU-FACS [2]	Manual	Lab	18-30	C	Yes	100	No	No	No
Semaine [13]	Manual	Lab	?	C	Yes	75	Yes	No	No
UT-Dallas [14]	Manual	Lab	18-25	C	Yes	284	No	No	No
VAM [6]	Manual	CTR	?	C	Yes	20	No	No	No

TABLE I

COMPARISON OF TEMPORAL FACIAL EXPRESSION DATABASES. C = CONTROLLED, CTN = CLOSE TO NATURAL AND CTR = CLOSE TO REAL.

on the WWW, the labelers 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, robust facial expression analysis methods have been developed, which along with their different databases have followed different 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.

AFEW is an acted facial expressions dataset in challenging conditions. Given the huge amount of video data on the WWW, it is worthwhile to investigate the problem of facial expression analysis in tough conditions. The video clips have been labelled for six basic expressions *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise* and the *neutral* class. The database captures varied facial expressions, natural head pose movements, occlusions, subjects from various races, gender, diverse ages and multiple subjects in a scene. In Section VI, we present the baseline results, which show that current facial expression recognition

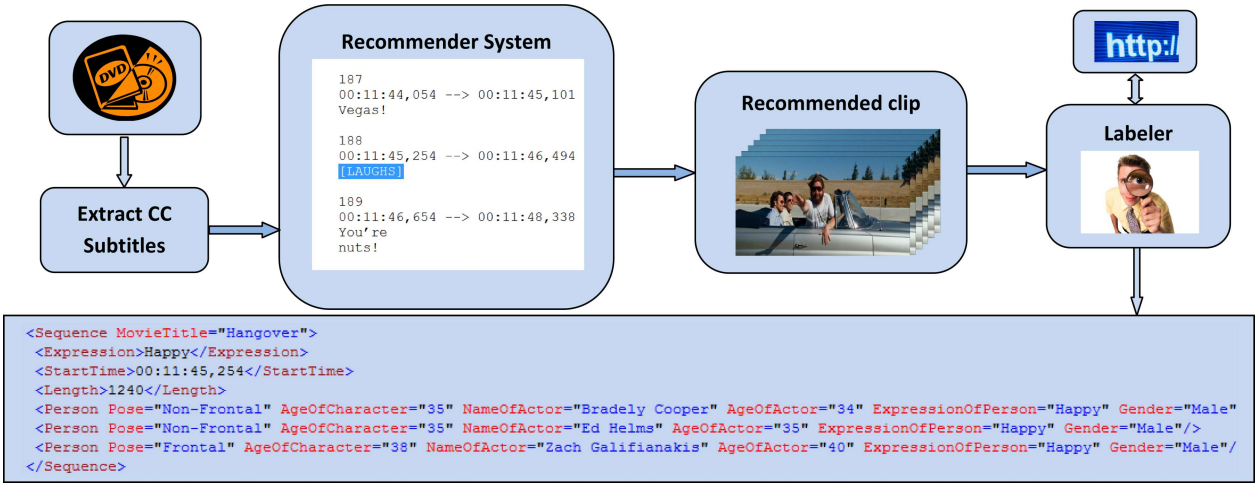

Fig. 1. The diagram describes the process of database creation. A subtitle is extracted from the DVD and then parsed by the recommender system. In the example here, when the subtitle contains the keyword 'LAUGH', the corresponding clip is played by the tool. The human labeler then annotates the subjects in the scene, using a GUI tool, based on the information about the subjects in the clip available on the WWW. The resulting annotation is stored in the XML schema shown at the bottom of the diagram. Note the structure of the information about a movie scene containing multiple subjects. The image in the diagram is from the movie 'The Hangover'.

approaches that have been reported to achieve high recognition rates on existing datasets do not cope with such realistic environments, underpinning the need for a new database and further research!

While movies are often shot in somewhat controlled environments, they provide significantly closer to real-world environments than current lab-recorded datasets. No claim is made that the AFEW database is a spontaneous facial expression database. However, clearly, (good) method actors attempt mimicking real-world human behaviour such that audiences get the illusion that they behave spontaneously, not posing, in movies. The 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.

A. *Related databases*

Over the past decade, many databases have been published. One of the earliest is the widely used Cohn-Kanade database [11]. It contains 97 subjects, which posed in a lab situation for the six universal expressions and the neutral expression. Its extension CK+ [11] contains 123 subjects but the new videos were shot in a similar environment. The Multi-PIE database [7] is a popular database and contains both

temporal and static samples recorded in the lab over five different sessions. It contains 337 subjects covering different pose and illumination scenes. The method of construction of these databases is purely manual where the subjects posed sequentially. The MMI [16] database is a searchable temporal database with 75 subjects. All of these are posed lab-controlled environment databases. The subjects display various acted (not spontaneous) expressions. The recording environment is nowhere close to real-world conditions.

The RU-FACS database [2] is a FACS-coded temporal database exhibiting spontaneous facial expressions, but it is proprietary and unavailable to other researchers. The Belfast database [5] consists of a combination of studio recordings and TV programme grabs labelled with particular expressions. The number of TV clips in this database is sparse. Compared to the manual method used to construct and annotate these databases, our recomender system makes the construction faster and 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 need to be searched manually.

The *JAFFE database* [12] is one of the earliest static facial expression datasets. It contains 219 images of 10 Japanese females. However, it has a limited number of samples, subjects and has been created in a lab controlled environment. Table I shows a detailed comparison of facial expression databases. In one of the first experiments on close to real data, [15] proposed a bimodal, audio-video features based system. The database has been constructed from TV programs. However, the size of database is fairly small, with 107 clips only.

Our AFEW database is similar in spirit to the Labeled Faces in the Wild (LFW) database [9] and the Hollywood Human Actions (*HOHA*) dataset [10]. These contain varied pose, illumination, age, gender and occlusion. However, LFW is a *static* face recognition database created from single face images found on the WWW specifically for face recognition and HOHA is an action recognition database created from movies.

II. Contributions of the database

Our databases have the following novelties:

- 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 seven facial expression classes.

- The subjects have a wide age range (1-70yr), which makes it very generic in terms of age, unlike other facial expression databases. The databases have a large number of 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, which has multiple labelled subjects in the same frame. This enables an interesting study on the 'theme' expression of a scene with multiple subjects, which may or may not have the same expression at a given time.
- The database exhibits 'close-to-real' illumination conditions. The clips stem from scenes with indoor, night-time and outdoor natural illumination. While it is clear that movie studios use controlled illumination conditions even in outdoor settings, we argue that these are closer to natural conditions than lab-controlled conditions and, therefore, 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 face recognition, face alignment, age analysis and action recognition.
- The movies have been chosen to cover a large set of actors. Many actors appear in multiple movies in the dataset, which will enable to study how their expressions have evolved over time, whether they differ for different genres, etc.
- 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 train and test splits for both temporal and static FER experiments.

III. Database creation

In this section, we discuss the details of the database construction. A semi-automatic approach was followed during the creation of the database. The process was divided into two parts. In the first step, the subtitles are extracted and parsed in the recommender system. In the second step, the labeler annotates the recommended clips based on the information available on the internet.

A. Subtitle extraction

We purchased and analysed fifty-four movie DVDs. We extracted the *Subtitles for Deaf and Hearing impaired (SDH)* as well as *Closed Caption (CC)* subtitles from the DVDs, as these types of subtitles

Attribute	Description
Length of sequences	300-5400 ms
Number of sequences	1426
Total number of expressions (incl. multiple subjects)	1747
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

TABLE II
ATTRIBUTES OF AFEW DATABASE.

contain information about the audio and non-audio context such as emotions, information about the actors and scene, e.g. '[CHEERING]', '[SHOUTS]', '[SURPRISED]', etc. We extracted the subtitles from the movies using the VSRip tool (http://www.videohelp.com/tools/VSRip). For the movies where VSRip could not extract subtitles, we downloaded the SDH subtitles from the WWW. The extracted subtitle images were parsed using Optical Character Recognition (OCR) and converted into .srt subtitle format using the Subtitle edit tool (www.nikse.dk/se). The .srt format contains the start time, end time and text content with milliseconds accuracy.

B. 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], [DISAPPOINT], [AMAZED] etc.). If found, the system recommends video clips to the labeler. The start and end time of the clip is extracted from the subtitle information. The system plays the video clips sequentially and the labeler enters information about the clip and its characters / actors from the WWW. In the case of clips with multiple actors, the sequence of labeling was 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 entered the age of the character

or, where this information was unavailable, estimated the age.

In total, the subtitles from the fifty-four DVDs contained 77666 individual subtitles. Out of these, the recommender system suggested 10327 clips corresponding to subtitles containing expressive keywords. The labelers chose 1426 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. Subtitles are manually created by humans and can contain errors. This may lead to a situation where the recommender system may suggest an erroneous clip. However, such a recommendation can be rejected by the labelers. The labelers annotated the clips based on the video, audio and subtitle information, so that they could make a more informed decision. The proposed recommender system can be used to easily add more clips to the database and scale it up to web scale.

C. 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 subjects in the clips.

- *Expression* - This specifies the theme expression conveyed by the scene. The expressions were divided into six expression classes plus neutral. The default value is based on the search keyword found in the subtitle text, for example for 'smile' and 'cheer' it is *Happiness*. The human observer can change it based on their observation of the audio and scene of the clip.
- *StartTime* - This denotes the start timestamp of the clip in the movie DVD and is in the hh:mm:ss,zzz format.
- *Length* - Duration of the clip in milliseconds.
- *Person* - This contains various attributes describing the actor / 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 non-frontal.
  - *AgeOfCharacter* - Where the age of the character was available from the WWW, this information was used. Frequently, this was only the case for the characters of the lead actors. Otherwise, the labeler estimated the age.
  - *NameOfActor* - Real name of the actor.
  - *AgeOfActor* - Real age of the actor. The labeler extracted the information from www.imdb.com. In a very few cases, the age information was missing, in which case the labeler estimated it.
  - *ExpressionOfPerson* - This denotes the expression class of the character as labelled by the human observer. This may be different from the 'Expression' tag as there may be multiple people in

the frame showing different expressions with respect to each other and the scene/theme.

- *Gender* - Gender of the actor.

This XML-based metadata schema has two major advantages. First and foremost, it is easy to use and search using any standard programming language on any platform that supports XML. Secondly, the structure makes it simple to add new attributes about the video clips, such as pose of the person in degrees and scene information, in the future, while keeping the existing data and ensuring that the already existing tools can take advantage of this information with minimal changes. Currently, the database metadata indexes 1426 video clips. Details of the database are in the Table II. The details on how to obtain the database and its experimental protocols will be made available at:

http://cs.anu.edu.au/few

D. Movies in the database

The fifty-four movies used in the database are: 21, About a boy, American History X, Black Swan, Bridesmaids, Change Up, December Boys, Did You Hear About the Morgans?, Dumb and Dumberer: When Harry Met Lloyd, Four Weddings and a Funeral, Friends with Benefits, Frost/Nixon, Ghoshtship, Hall Pass, Halloween, Halloween Resurrection,Harry Potter and the Philosopher's Stone, Harry Potter and the Chamber of Secrets, Harry Potter and the Deathly Hallows Part 1, Harry Potter and the Deathly Hallows Part 2, Harry Potter and the Goblet of Fire, Harry Potter and the Half Blood Prince, Harry Potter and the Order Of Phoenix, Harry Potter and the Prisoners Of Azkaban, It's Complicated, I Think I Love My Wife, Jennifer's Body, Little Manhattan, Messengers, Notting Hill, One Flew Over the Cuckoo's Nest, Orange and Sunshine, Pretty in Pink, Pretty Woman, Remember Me, Runaway Bride, Saw 3D, Serendipity, Something Borrowed, Terms of Endearment, The American, The Aviator, The Devil Wears Prada, The Hangover, The Haunting of Molly Hartley, The Informant!, The King's Speech, The Pink Panther 2, The Social Network, The Terminal, The Town, Valentine Day, Unstoppable, You've Got Mail.

E. SFEW

Static facial expression analysis databases such as Multi-PIE and JAFFE [12] are lab-recorded databases in tightly controlled environments. For creating a static image database that represents the real world more closely, we extracted frames from AFEW. This static database is named *Static Facial Expressions in the Wild (SFEW)*. The *Strictly Person Independent* version of SFEW is described in [4] and is posted as a challenge on the BEFIT website (http://fipa.cs.kit.edu/511.php). In Section V, we describe the three

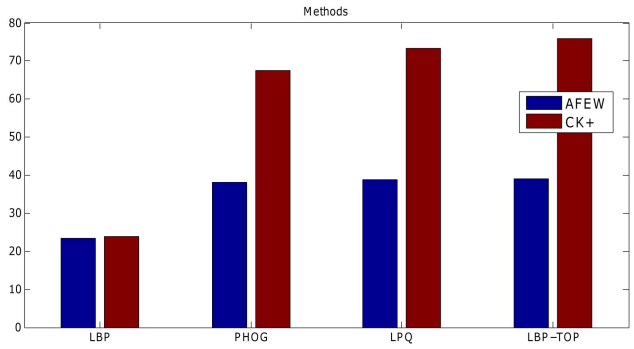

Fig. 2. The graph shows the performance of LBP, PHOG, LPQ and LBP-TOP on the CK+ and DFEW databases.

versions of SFEW, which are based on the level of subject dependency for evaluating facial expression recognition performance of systems in different scenarios.

IV. Comparison with other databases

In this section, we compare the performance of state-of-the-art descriptors on AFEW and SFEW with that on existing widely used datasets. AFEW is compared to the CK+ database, which was introduced by Lucey *et al.* [11] as an extension of the Cohn-Kanade database. A basic facial expression comprises of various temporal dynamic stages: *onset*, *apex* and *offset* stage. In CK+, all videos follow the temporal dynamic sequence: *neutral* $\rightarrow$ *onset* $\rightarrow$ *apex*, which is not a true reflection of how expressions are displayed in real-world situations as the data about the offset phase is missing. We also argue that the complete data containing the complete temporal sequence may not be always available. For example: a person entering a scene may already be happy and close to the highest intensity of happiness (onset). Earlier systems trained on existing databases like CK+ have learnt on the above mentioned 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. For extracting the face, we computed the Viola-Jones detector [17] 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: (1) a comparison of SFEW, JAFFE and Multi-PIE on the basis of four common expression classes (*disgust*, *neutral*, *happiness* and *surprise*) and (2) a comparison of SFEW and JAFFE on seven expression classes.

We computed feature descriptors on the cropped faces from all databases. The cropped faces were divided into 4×4 blocks for LBP [8], LPQ [8] and PHOG [3]. For LBP and LPQ, the neighbourhood

Protocol	**AFEW/SFEW (Train-Test sets contain:)**
Strictly Person Specific (SPS)	same single subject
Partial Person Independent (PPI)	a mix of common and different subjects
Strictly Person Independent (SPI)	different subjects [4]

TABLE III

EXPERIMENTATION PROTOCOL SCENARIOS FOR SFEW AND AFEW.

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%	33.8%
SFEW_SPI	23.0%	13.0%	13.9%	29.0%	23.0%	17.0%	13.5%	18.9%
SFEW_SPS	35%	-	45.8%	0.0%	7.1%	-	0.0%	17.5%

TABLE IV

AVERAGE CLASSIFICATION ACCURACIES OF DIFFERENT PROTOCOLS.

size was set to 8. For PHOG, bin length = 8, pyramid levels L=2 and angle range = [0,360]. We applied principal component analysis (PCA) on the extracted features and kept 98% of the variance. For classification, we used a support vector machine learned model. The type of kernel was C-SVC, with a radial basis function (RBF) kernel. We used five-fold cross validation for selecting the parameters. For AFEW, the static descriptors were concatenated. LBP-TOP performs best out of all the methods. For all the methods, the overall expression classification accuracy is much higher for CK+ (see Figure 2).

For SFEW's ***four expression class*** experiment, the classification accuracy on the Multi-PIE subset is 86.25% and 88.25% for LPQ and PHOG, respectively. For JAFFE, it is 83.33% for LPQ and 90.83% for PHOG. For SFEW, it is 53.07% for LPQ and 57.18% PHOG. For the ***seven expression class*** experiment, the classification accuracy for JAFFE is 69.01% for LPQ and 86.38% for PHOG. For SFEW, it is 43.71% for LPQ and 46.28% for PHOG. It is evident that LPQ and PHOG achieve 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 all these state-of-the-art methods have been experimented

on are recorded in lab-based environments. Expression analysis in (close to) real-world situations is a non-trivial task and requires more sophisticated methods at all stages of the approach, such as robust face localisation/tracking, illumination and pose invariance.

V. Experimentation Protocols

Over the years, many facial expression recognition methods have been proposed 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, as defined in Table III.

The BEFIT workshop challenge [4] falls under SPI for SFEW. Data, labels and other protocols will be made available on the database website. AFEW_PPI contains 745 videos and AFEW_SPI contains 741 videos in two sets. AFEW_SPS contains 40 videos of Daniel Radcliffe for four expression category (*fear*, *happiness*, *neutral* and *surprise*). For SFEW, SFEW_SPS contains 76 images of the actor 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. In the next section, we describe the baselines for all these protocols, with the data divided into two sets. The results should be reported as an average of training and testing on the sets.

VI. Baseline

For all the protocols for SFEW, we computed the baselines based on the method defined in [4]. PHOG [3] and LPQ [8] features were computed on the cropped face. The features were concatenated together to form a feature vector. For dimensionality reduction, PCA was computed and 98% of the variance kept. Further, a non-linear SVM was used to learn and classify expressions. See [4] for the parameter selection details. For AFEW for encoding the temporal data, we computed LBP-TOP [8] features as in Section IV. The classification accuracy for both databases and their protocols can be found in Table IV. Low classification accuracy results show us that the current methods are not appropriate for real-world scenarios.

VII. Conclusions

Collecting richly annotated, large datasets representing real-world conditions is a non-trivial task. To address this problem, we have collected two new facial expression databases derived from movies via a semi-automatic recommender based method. The database contains videos showing natural head poses

and movements, close to real-world illumination, multiple subjects in the same frame, a large age range, occlusions and searchable metadata. The datasets also cover toddler, child and teenager subjects, which are missing in other currently available temporal facial expression databases. AFEW also contains clips with multiple subjects exhibiting similar or different expressions with respect to the scene and each other. This will enable research on the effect of context/scene on human facial expressions. We compared our AFEW database with the CK+ database using state-of-the-art descriptors and SFEW with the Multi-PIE and JAFFE databases. We believe that these datasets will enable novel contributions to the advancement of facial expression research and act as a benchmark for experimental validation of facial expression analysis algorithms in real-world environments.

REFERENCES

[1] T. Bänziger and K. Scherer, "Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) Corpus," in *Blueprint for affective computing: A sourcebook*, K. Scherer, T. Bänziger, and E. Roesch, Eds. Oxford, England: Oxford University Press, 2010. 3

[2] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, 2006. 3, 5

[3] A. Bosch, A. Zisserman, and X. Munoz, "Representing Shape with a Spatial Pyramid Kernel," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR*, 2007, pp. 401–408. 10, 12

[4] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static Facial Expression Analysis in Tough Conditions: Data, Evaluation Protocol and Benchmark," in *Proceedings of BeFIT'11, IEEE International Conference on Computer Vision Workshop*, 2011, pp. 2106–2112. 2, 9, 11, 12

[5] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A New Emotion Database: Considerations, Sources and Scope," in *ISCA ITRW on Speech and Emotion*, 2000, pp. 39–44. 3, 5

[6] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *IEEE International Confernce on Multimedia and Expo ICME'08*, 2008, pp. 865–868. 3

[7] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *Proceedings of the Eighth IEEE International Conference on Automatic Face and Gesture Recognition FG'2008*, 2008, pp. 1–8. 3, 4

[8] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, pp. 1 –17, 2011. 10, 12

[9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, 2007. 2, 5

[10] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR'08*, 2008, pp. 1–8. 5

[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops CVPR4HB'10*, 2010, pp. 94–101. 3, 4, 10

[12] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets," in *Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition and Workshops FG'98*, 1998. 5, 9

[13] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Proceedings of the IEEE International Conference on Multimedia and Expo ICME'10*, 2010, pp. 1079–1084. 3

[14] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi, "A video database of moving faces and people," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 812–816, 2005. 3

[15] M. Paleari, R. Chellali, and B. Huet, "Bimodal emotion recognition," in *Proceeding of the Second International Conference on Social Robotics ICSR'10*, 2010, pp. 305–314. 3, 5

[16] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proceedings of the IEEE International Conference on Multimedia and Expo ICME'05*, 2005, pp. 317–321. 3, 5

[17] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, ser. CVPR'01, 2001, pp. 511–518. 10

[18] F. Wallhoff, "Facial Expressions and Emotion Database," 2006, http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html. 3

Abhinav Dhall is a PhD candidate at the Research School of Computer Science at the Australian National University. He received his bachelor's degree in computer science in 2006 from the DAV Institute of Engineering and Technology (India). He was awarded the Australian Leadership Award Scholarship 2010. His research interests include affective computing, computer vision, pattern recognition, and HCI.

Roland Goecke leads the Vision and Sensing Group, Faculty of Information Sciences and Engineering, University of Canberra. He received a Masters in Computer Science from the University of Rostock, Germany, in 1998 and his Ph.D. in Computer Science from the Australian National University in 2004. His research interests are in affective computing, computer vision, human-computer interaction and multimodal signal processing.

Simon Lucey is Senior Research Scientist and Research Project Leader at the CSIRO ICT Centre Computer Vision laboratory, Brisbane. He was awarded an ARC Future Fellowship by the Australian Research Council in 2009. He received his Ph.D. in Computer Science from the Queensland University of Technology, Brisbane, in 2002. His research interests are in pattern recognition, computer vision and machine learning.

Tom Gedeon is Chair Professor of Computer Science at the Australian National University, and President of the Computing Research and Education Association of Australasia. His BSc and PhD are from the University of Western Australia. He is a former president of the Asia-Pacific Neural Network Assembly, and serves on journal advisory boards as member or editor.