BBC

Research & Development

White Paper

WHP 231

September 2012

A Pilot Study for Mood-based Classification of TV Programmes

Jana Eggink, Penelope Allen, Denise Bland

BRITISH BROADCASTING CORPORATION

White Paper WHP 231

A Pilot Study for Mood-based Classification of TV Programmes

Jana Eggink, Penelope Allen, Denise Bland

Abstract

We report results from a pilot study on mood-based classification of TV programmes. Short video clips from various programmes were labelled on three mood axes, giving the subjects opinion of how happy, serious and exciting each clip was. This data was used for mood classification based on automatically extracted audio and video features in a machine learning framework. Attention was given to the challenges of dealing with a small dataset as commonly obtained from pilot studies, showing that a thorough evaluation was possible and produced useful results. Introducing a new feature based on face detection and combining it with other signal processing features led to good classification accuracies. These lay between 85% and 100% for the most simple setting, and still reached more than 70% accuracy when a finer three point mood scale was used. Overall, the results were promising and showed that automatic mood classification of video material is possible. Moods can therefore be used as additional metadata to facilitate search in large archives.

http://dl.acm.org/citation.cfm?id=2245276.2245455

Additional key words: Multimedia classification, mood classification, machine learning

White Papers are distributed freely on request.

Authorisation of the Chief Scientist or General Manager is required for publication.

White Paper WHP 231

A Pilot Study for Mood-based Classification of TV Programmes

Jana Eggink, Penelope Allen, Denise Bland

1 Introduction

Many professional media organisations, such as broadcasters, own extremely large multimedia archives. Making these archives searchable is a daunting task, as often only limited metadata exists. In the case of the BBC television archives, manually generated genre information for most programmes exists and is of course very valuable. However, even selecting a specific genre still gives too many results for comfortable searching and exploring by non-professional users. To open up the BBC archives to the general public as is currently planned, additional metadata is required. In this paper we explore the possibility of using mood as additional information, which can then be used for searching and browsing. Moods can also be linked to personal preferences and are therefore useful for recommendations.

For large archives, mood metadata needs to be computed automatically, as manual categorization is far too expensive. We compute low level audio and video features based on signal processing, and map them to high level semantic mood categories using machine learning techniques. We deal with restrictions imposed by a limited amount of human subject data from our pilot study and show meaningful results.

2 Literature Review

A number of published papers exist that are concerned with the problem of video classification. Most of these try to automatically infer the genre, with only a very small number attempting mood based classification. For genre, both audio and video features have been shown to be useful. Most commonly used as audio features are zero-crossing rate, sound energy, bandwidth, spectral centroid and Mel-Frequency Cepstral Coefficients (MFCCs). For video features colour, motion and shot-based features such as average shot length are most common. Due to the lack of a common data set, it is nearly impossible to say which features or machine learning approaches give best results. Nearly all publications use different videos and different classes, varying both in the definition and the number of classes. For an overview of existing approaches to video genre classification see [1].

One of the few publications dealing with mood-based classification of video is [5]. The authors were working with a psychological model based on arousal, valence and control (AVC-scale), but came to the conclusion that the control dimension is of limited relevance for video and excluded it. The arousal dimension is closely related to the feeling of energy, ranging from calm to excited. Valence describes the affective evaluation, ranging from pleasant and positive to unpleasant and negative. These two dimensions are modelled independently and with different feature sets, using motion, shot length and sound energy for arousal, while valence was based solely on estimated pitch during speech segments. They did not attempt to classify videos into distinct classes, but rather mapped them on continuous scales, varying over time as the video changed. While they indicated promising results on a single, specially selected movie, no formal or larger scale evaluation was carried out.

In a similar setting, [6] worked on detecting three different emotions (fear, sadness, joy) in movie scenes, using video features based on colour, motion and shot cut rate. For the classification a Hidden Markov Model (HMM) was used, which could also learn the transition likelihoods between different mood states. Classification accuracy for the best configuration was promising with numbers reaching 80%, but only three movies were included in the test set.

Another paper by [12] was also concerned with emotion detection in movies, mainly utilizing colour information. They used 16 mood terms, a much larger selection than previous studies, and 15 movies of different genres. Reported accuracy was also around 80%, but the use of and results for individual mood terms remained somewhat unclear.

3 Data Collection

The published results for mood based classification of video material are limited, and no common dataset exists. Most work concentrated on movies. The BBC archives however contain material from very different genres, ranging from soaps and drama to news and documentaries, so a purpose made data collection was necessary. However, conducting a large scale user trial is expensive, and we first wanted to establish that our approach had the potential to lead to useful results.

The ground truth used in this paper is based on a small user study, which investigated programme mood perception and categorization. Participants assigned mood scores to a limited number of TV programmes they were engaged with, these served as ground truth for the classification results in this paper. Of the nine participants, four were female, five were male, ranging in age from 25 to 60 years.

Mood ratings were assigned on a sematic differential scale based on the Evaluation, Potency and Activity (EPA) structure developed by [9]. In a study on a large collection of semantic differentials, Osgood found three reoccurring dimensions that constitute most affective meaning. The three elements together create a semantic space in which similar stimuli can be compared. In the EPA framework, *E* for evaluation accounts for about 50% of all affective meaning, it is related to how positive we assess something to be*. E* includes differentials such as happy-sad or light-dark. *P* for potency is related to powerfulness which includes differentials such as serious-funny, masculine-feminine and *A* denotes activity, for example exciting-relaxing or dramatic-calm. These differentials were the scales used in the pilot study.

The EPA scale is very similar to the AVC scale used in some literature [5], our choice to use EPA is that it has been tested and is applicable cross culturally. Participants were required to watch a three minute exert from a TV programme and rate the clips moods on the related EPA differential scales. These scales were five point Likert scales [7] spanning the space between the opposing adjectives. Participants were then asked to expand and give reasoning for their evaluation.

Ideally, the first step would look at user agreement, to find out if the mood classification was based on some commonly agreed ‘truth’, or if it was mainly influenced by participants individual differences and therefore carries high variance. In the trial, one programme was assessed by more than one participant (from the UK TV series Eastenders). Agreement in mood evaluation for this programme was high; all five ratings were skewed towards happy/light on the evaluation scale, and all but one rating skewed toward exciting/dramatic on the activity scale. For the potency scale, all but one participant rated 3, expressing either a neutral or combined assessment on the scale. This finding may be down to investigator effects from the masculine/feminine scale. Anecdotal feedback from the participants revealed that they weren’t comfortable expressing an opinion with these adjectives. The original research for the EPA scale was published in the late 1950’s, and society has undeniably changed since then, especially with respect to gender issues. While a more detailed analysis of this issue might be interesting, it is outside the scope of this paper. Even though we couldn’t reliably estimate user agreement from one programme only, we assumed that if we could successfully build a user independent classifier, there must be some underlying agreement that the classifier was able to learn.

4 Classification Set-up

For the purpose of automatic classification, to maximize our use of the limited data, we opted for a leave-one-out cross validation strategy, training on all but one example, which was then used for testing. This was repeated until all video clips have been the test example once. The same classifier parameters were used for all test examples to ensure the classifier is able to generalize. The average results of a full round of cross validation are reported.

We have three axes, Evaluation (E), Potency (P), and Activity (A), each with a five point scale, in theory this means we have 15 classes. Ideally, we would make use of the ordinal (rank-ordered) nature of the data, where both a rating of four and one of five are skewed toward happy/light on the evaluation axis, but five is expressing a stronger opinion. However, due to the limited amount of data obtained in the pilot study, we decided for a simpler set-up where each class was treated independently, even if this meant we might not make use of all available information. We assume that the absolute minimum of examples for a usable class is three. If we'd have only one example for training, it would be extremely difficult for any algorithm to generalise. To obtain realistic variation in the data, we required at least two examples for training, and one more for testing.

As a result of the selection, we had sufficient examples for classes 3, 4, and 5 on the evaluation axis; 2, 3, and 4 on the potency axis; and 3 and 4 on the activity axis. For both the evaluation and potency axis there were 17 usable examples (out of 18 originally labelled by participants), and 13 for activity. For a distribution of the individual video clips in the EPA space see Figure 1.

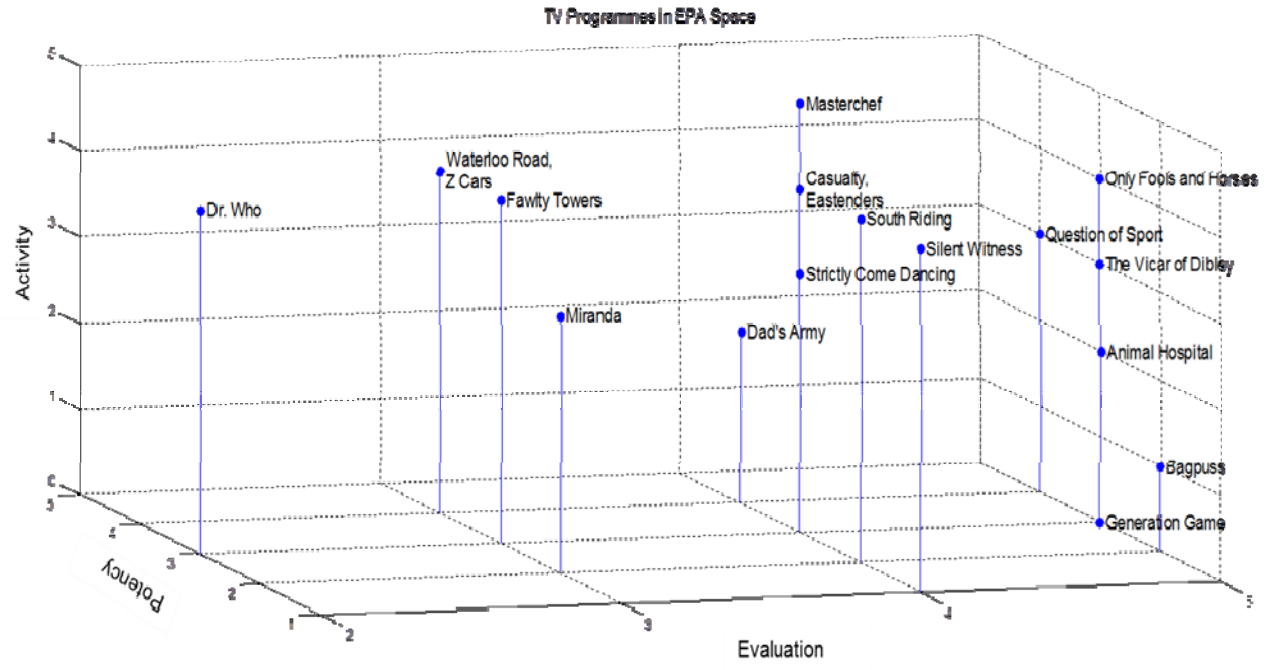

Figure 1. Overview of TV Programme Clips in EPA space

5 Features and Classifier

Features from both the audio and video were extracted using signal processing. These features were then used to train a classifier, mapping them to the higher-level mood. Only features from the three minute clips were used, corresponding to the same video material that was watched and rated by the participants.

For the audio we used standard features, Mel-Frequency Cepstral Coefficients (MFCCs), a spectral representation with a frequency resolution similar to human hearing. The first coefficient, commonly called C0, codes the overall energy at that point in time. We also used MFCC deltas, coding the short term change or variability. For details and software used see [3]. Parameters in this study were: The first 20 MFCCs or C0 on its own, based on a frame length of 0.025 seconds with no overlap between frames, a frequency range from 50Hz to 8000Hz and a delta range of ±10 frames.

Our video features were based on the presence of faces, average luminance, motion and cut frequency. While colour has been used [12], we decided against it, because part of the archive is very old and therefore in black and white only. The extraction of all video features except the face-based ones were from down sampled tiny video images [11]. A tiny video image is a 32 by 32 pixel grey scaled version of a standard definition resolution coloured frame. Each video frame image is

resized to 32 by 32 pixels and converted to grey scale from a weighted sum of the R, G and B components.

$$\text{Grey scale} = (0.2989 * R) + (0.5870 * G) + (0.1140 * B)$$

The summation luminance value LUM for frame k is the pixel summation of the tiny grey image.

$$LUM = \sum\sum f_k(\text{pixel})$$

The cuts feature value indicates a straight cut between adjacent frames. It is produced from the threshold product of the mean absolute difference [4] multiplied by one minus the phase correlation [10] where the phase correlation is the difference between the current tiny grey image f_k and previous tiny grey image f_{k-1}.

The mean absolute difference (MAD) follows, where S_1 is scaling:

$$MAD = S_1 \sum\sum |f_k - f_{k-1}|$$

The phase correlation (PHC) follows, where F_k is the Discrete Fourier Transform (DFT) of f_k and F*k is the complex conjugate of the DFT of f_k:

$$\text{PHC} = \text{maximum of } \text{DFT}^{-1} \left(\frac{F_k F^*_{k-1}}{\left| F_k F^*_{k-1} \right|} \right)$$

The cuts feature value (CFV) is:

$$CFV = (MAD*(1-PHC)) > \text{threshold}$$

The amount of motion between two frames was taken as the difference between non adjacent tiny grey images. It was calculated from the difference between tiny grey images f_k separated by 10 frames (0.4 seconds) to give a smooth motion. The motion feature was produced from the mean absolute difference between the current frame's grey image f_k and the tenth previous frame's grey image f_{k-10} minus any cuts CFV. Deducting cuts blanked the motion feature for 10 frames.

The motion feature value (MFV) follows, where S_2 is scaling:

$$MFV = S_2 \sum\sum |f_k - f_{k-10}| - CFV_k$$

A measure of the number of continuous, long duration full frontal faces was also used. Firstly, the OpenCV face detector [8] was used to extract faces. Then for each frame, the presence or absence of a face was recorded and the cumulative sum of detected faces was calculated on a frame by frame basis and reset to zero whenever no face was detected. The face feature value was the cumulative sum of faces scaled by the face diameter. The motivation for this feature was based on personal observation. We noticed that faces present for a continuous time, for example a person speaking directly to the camera, coincided with more serious programmes.

Audio and video features were computed at different frame rates. Where necessary, alignment was performed by locally averaging the audio features to match the lower frame rate of the video.

Support Vector Machines (SVMs) were chosen as classifiers, using the implementation of libSVM [2]. Only radial basis function (RBF) kernels were used, with the two main parameters (C, controlling the punishment of misclassification of individual examples during training, and gamma, the kernel spread influencing generalization abilities) optimized in a full grid search separately for each feature combination. The same parameters were used for all files within one run of leave-one-out cross-validation. In case of more than 2 classes, we used the libSVM inbuilt extension to multi-class problems based on multiple binary classifiers.

Final classification was either based on individual frames, or on a per file average of all features. In the latter case, we used both mean and standard deviation of each feature type. Frame-based classification is computationally more expensive, but has the advantage of tracking changes in mood over time within a video programme.

6 Classification Results

During testing we obtained a classification result either for each file, or for each frame of a file. We only had one label for the entire file and no time-varying or frame specific labels, so we assumed that the label is valid for all individual frames. In cases where we obtained frame-based classification results but wanted to average them, we selected the most frequent class for each file. Other heuristics, including some that give different weight or confidence to different frames, would be possible but were not evaluated.

To interpret any error measure a baseline was needed. In the simplest case we have a two class problem, where each test item is either class A or class B. Random guessing would result in a 50% success rate. This is certainly justified if the two classes are evenly distributed, which in many real-world problems is not the case. If the relative class distribution is 80% to 20%, a baseline on always assuming the most frequent class is classifying 80% of cases correct. A machine learning based algorithm is only useful when it is better than the baseline. Any reported error measure is therefore only informative in combination with an appropriate baseline, or at least information about the relative distribution of classes. In the following, a baseline measure was based on always selecting the most frequent class.

6.1 Feature Selection

The first question to answer was which features were useful for classification. In theory, if we had sufficient training examples, the classifier should learn that automatically. As we had a very limited number of examples, it was likely that features without useful information would just add noise and degrade the classification results. We were testing both audio and video features, assuming they contain complementary information which together would improve classification.

For a basic analysis of the features, we started with the simplest set-up, using averaged features values for each file, and only two classes on each axis, being as different as possible given the available data.

For the evaluation axis, we used classes 3 against 5 (neutral/combined against strongly skewed towards happy/light); for potency we used 2 against 4 (skewed towards funny/feminine against skewed towards serious/masculine); and for activity 3 against 4 (neutral/combined against skewed towards dramatic/exciting).

Figure 2 shows the number of correctly classified videos given different feature combinations. It also shows the baseline obtained when always choosing the most frequent class. For the activity scale, a 70% baseline is quite high due to the uneven distribution of labels. Using for example cuts or luminance as sole features also gave 70% correct classification, but as this was exactly the baseline result, these features did not appear to be useful.

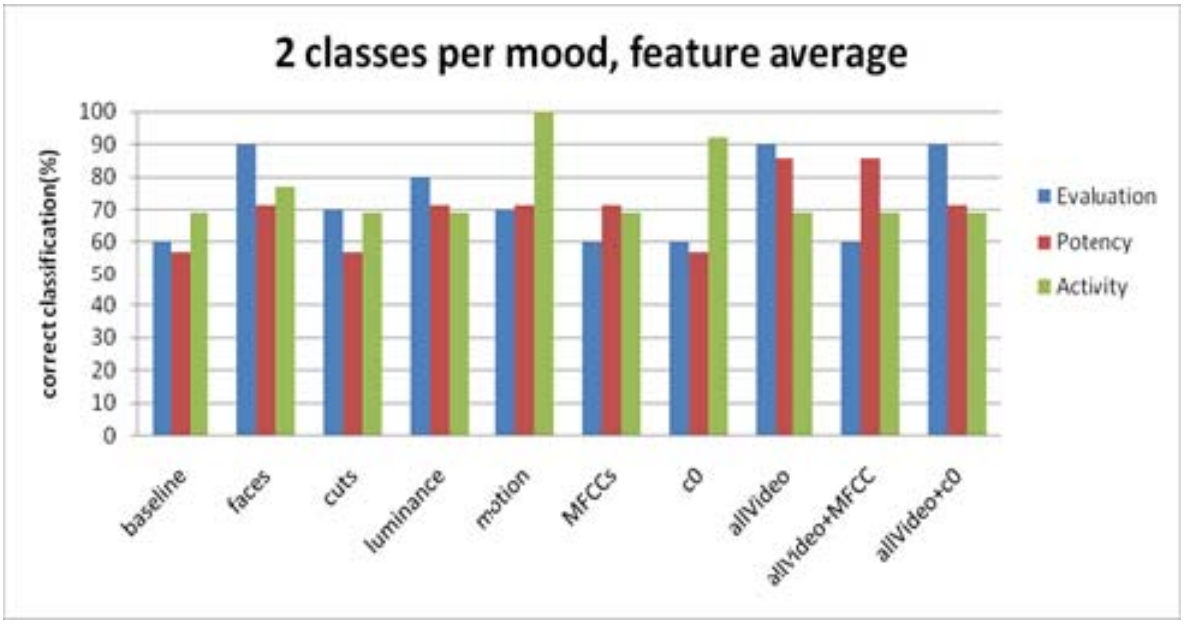

Figure 2. Classification Results for individual Features,
2 classes per mood axis, features averaged per file

It can be clearly seen that the classification of evaluation, potency and activity required different features. For evaluation, the single best feature was the face-based one, with luminance in second place. The audio-based MFCCs are at baseline. Adding MFCCs to the video features actually degraded the results to baseline performance, probably because the higher number of MFCCs overshadowed the useful information in the 4 video features. For potency there seems no clear advantage of specific features, and a combination of all video features, potentially joined by MFCCs, seemed most promising. The activity axis seemed to be best modelled by the motion feature. C0, being related to loudness also gave good results, with the face-based feature the only other feature performing above baseline. For the remaining experiments, two promising feature combinations were chosen for each axis, see Table 1.

	Combination 1	Combination 2
Evaluation	faces + luminance	faces + luminance + C0
Potency	all video features	all video features + MFCCs
Activity	motion + C0	motion + faces + C0

Table 1. Feature Selection

6.2 Three Classes per Mood

In addition to the feature selection, we also expanded the number of classes used for each mood axis, resulting in three classes for evaluation and potency, and two for activity, based on the selection criteria described in Section 4. Results are shown in Figure 3, again using features averaged per video clip. Classification accuracy for evaluation is 76%, nearly twice as good as the baseline of picking the most frequent class. Adding audio was not useful. For potency the results are only around 10% above baseline when using all video features, again adding audio degraded accuracy. Activity has only two usable classes, and results are very good with 92% or 100% accuracy above a baseline of 69%.

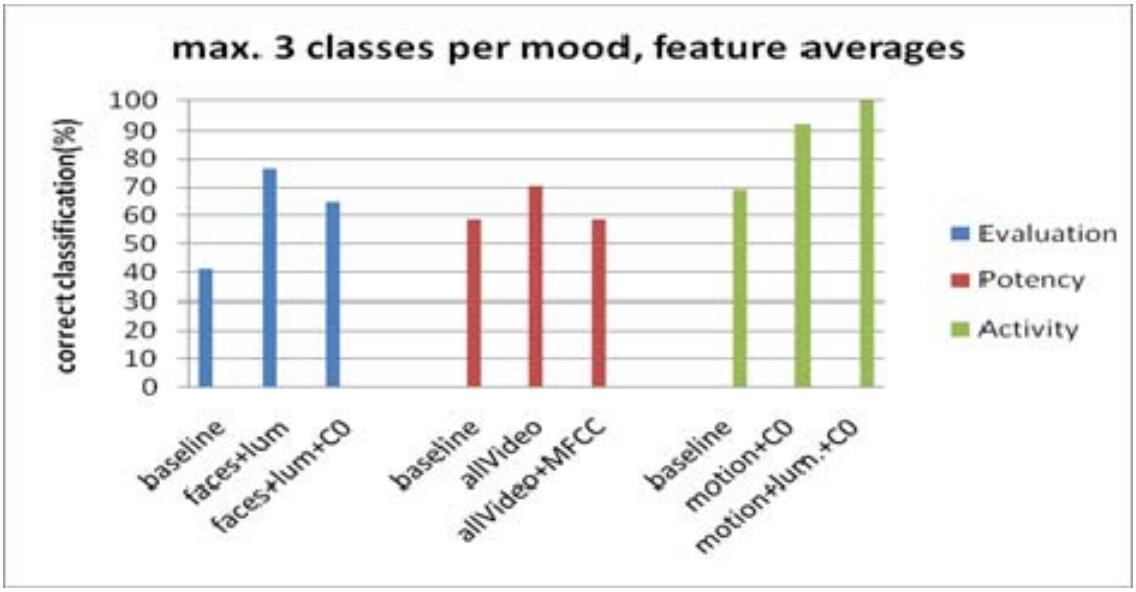

Figure 3. Classification results for selected features,
max. 3 classes per mood axis, features averaged per file

6.3 Frame-based Classification

For the next experiment we changed to frame-based classification. We had human labels for entire files, and didn't know if or how much perception might have changed during the three minute clips, but wanted to test if frame-based classification is possible. To compute the final accuracy, we either counted the number of correctly classified individual frames, or computed a result for each video file first, and then averaged these. For the latter, we took the most frequent class from the frame-based classification to be the overall classification results for a file (called 'file' below). Results for both evaluations are shown in Figure 4, using two classes per mood axis. Accuracy for the evaluation scale was still quite high when results were averaged for each file, but relatively low in terms of individual frames correctly classified. Adding audio information in the form of C0 improved the file averaged version only, indicating that it only led to a more favourable distribution of correctly classified frames, without improving overall reliablity. For potency, adding MFCCs degraded results, similar to the previous condition were feature averages where used. The strongest drop in performance occoured for the activity scale. This might be caused because it is much harder to cover the variance of a feature if each frame is classified individually, but this would need further investigation.

As a last step, we used frame-based classification and the available three classes for the evaluation and the potency axis. For the evaluation axis, best results were obtained when using faces, luminance and C0 as features. When classification results were averaged for each file, 53% were correctly classified, against a baseline result of 41%. For potency, results were at baseline only. This was not totally unexpected, as results for the averaged features were also low. It was probably caused by the problems with the chosen adjective pair masculine-feminine, and the resulting neutral rates from the subjects.

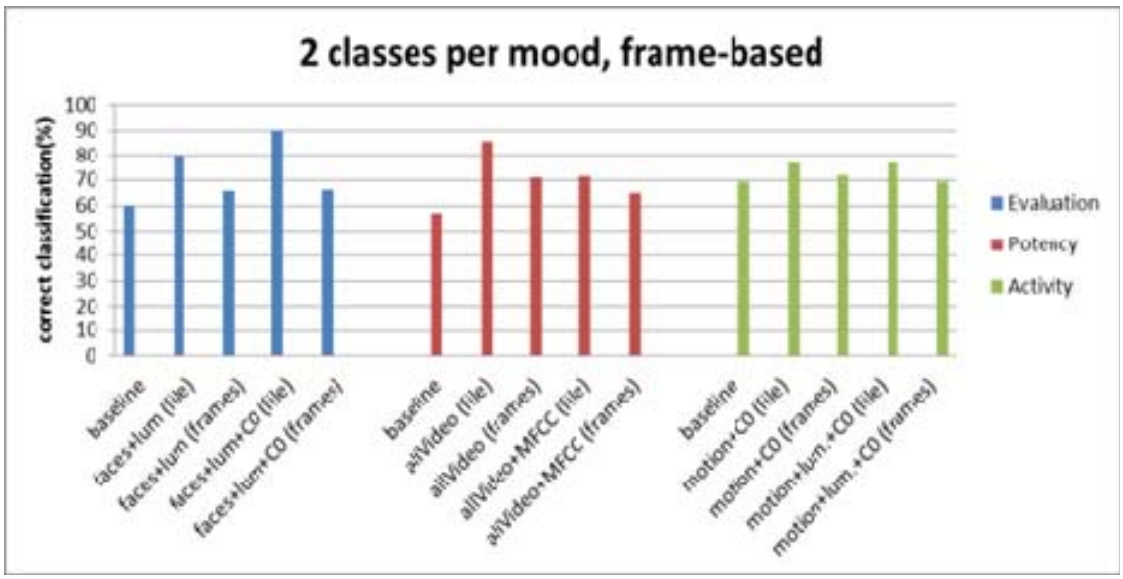

Figure 4. Classification Results for individual Features,
2 classes per mood axis, frame-based

7 Conclusions and Future Work

Overall, classification results were very encouraging, and the new feature based on the continuous presence of a face also proved to be useful for specific moods. Classification based on features averaged over a three minute clip gave accuracies between 85% and 100% when using two classes per mood axis, and between 70% and 76% when a finer scale with three classes per mood axis was used.

Using a cross validation scheme that maximises the use of available data and looking at the data from multiple perspectives was a useful approach to gain confidence in results obtained from a relatively small dataset. The high accuracy for the evaluation and the activity mood axes also indicated that there was some agreement between participants; if it was an entire subjective impression the classifier could not have learnt the labels assigned by different participants. The potency mood axis seems less useful, at least with the adjectives used in this study. This had already been indicated by the feedback from the participants, and could again be seen in the low classification results.

While there remains plenty of work to be done, and we are currently carrying out a large scale data collection, the pilot study was very useful. It confirmed our decision for an investment in a large study, and helped to build a working prototype needed for further user studies.

8 References

[1] Brezeale, D., and Cook, D.J. Automatic Video Classification: A Survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics*, 38 (3), 2008.

[2] Chang, C.-C. and Lin, C.-J. LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27*, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[3] Ellis, D.P.W. *PLP and RASTA (and MFCC, and inversion) in Matlab*. http://www.ee.columbia.edu/~dpwe/ resources/matlab/ rastamat/, 2005

[4] Easterbrook, J. *Source code. source forge link*, 2002. https://sourceforge.net/projects/shot-change/.

[5] Hanjalic, A. and Xu, L.-Q. Affective Video Content Representation and Modeling. *IEEE Transactions on Multimedia, 7 (1),* 2005

[6] Kang, H.-B., Affective content detection using HMMs. *Proceedings of the eleventh ACM international conference on Multimedia*, 2003

[7] Likert, R. A Technique for the Measurement of Attitudes. *Archives of Psychology 140: 1–55,* 1932

[8] *Opencv face detection.* http://opencv.willowgarage.com/wiki/FaceDetection.

[9] Osgood, C.E., Suci, G. and Tannenbaum, P. *The Measurement of Meaning.* University of Illinois Press, 1957

[10] Ren, J., Jiang, J., and Chen, J. Determination of shot boundary in mpeg video for trecvid 2007. *TRECVID 2007 workshop participants notebook papers*, 2007.

[11] Torralba, A., Fergus, R. and Freeman, W. T. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(11):1958–1970*, 2008.

[12] Wei, C.-Y., Dimitrova, N., and Chang, S.-F. Color-Mood Analysis of Films Based on Syntactic and Psychological Models. *IEEE International Conference on Multimedia and Expo,* 2004.