THE QUEEN'S AWARDS
FOR ENTERPRISE:
INNOVATION
2011

BBC

Research & Development

White Paper

WHP 232

September 2012

A Large Scale Experiment for Mood-based Classification of TV Programmes

Jana Eggink, Denise Bland

BRITISH BROADCASTING CORPORATION

White Paper WHP 232

A Large Scale Experiment for Mood-based Classification of TV Programmes

Jana Eggink, Denise Bland

Abstract

We present results from a large study with 200 participants who watched short excerpts from TV programmes and assigned mood labels. The agreement between labellers was evaluated, showing that an overall consensus exists. Multiple mood terms could be reduced to two principal dimensions, the first relating to the seriousness or light-heartedness of programmes, the second describing the perceived pace. Automatic classification of both mood dimensions was possible to a high degree of accuracy, reaching more than 95% for programmes with very clear moods. The influence of existing human generated genre labels was evaluated, showing that they were closely related to the first mood dimension and helped to distinguish serious form humorous programmes. The pace of programmes however could be more accurately classified when features based on audio and video signal processing were used.

This document was originally published in the Proceedings of the IEEE International Conference of Multimedia and Expo (ICME), Melbourne, Australia, July 2012. © IEEE

Additional key words: Multimedia classification, mood, genre, signal processing, machine learning, human labelling

White Paper WHP 232

A Large Scale Experiment for Mood-based Classification of TV Programmes

Jana Eggink, Denise Bland

1 Introduction

The British Broadcasting Corporation (BBC) is one of the largest public broadcasters, with archives reaching back to the very beginnings of radio and television. To open up large-scale multimedia archives and make them accessible to non-professional users, metadata is required. In the special case of the BBC archives, some manually created metadata is available, including genre information for most TV programmes. However, this information is not always sufficient for effective searching and browsing, and we explore the possibility of using mood as additional metadata. Manual labelling is far too costly for any large archive, and we therefore focus on using automatically extracted features and machine learning techniques. We also investigate the use of existing human generated genre labels for mood classification.

2 Literature Review

Most publications in the area of video classification are concerned with genre rather than mood. For genre, both audio and video-based signal processing features have been shown to be useful, for an overview see [1].

One publication aiming at mood-based classification of video is [5]. The authors worked with two independent mood dimensions, valence (affective evaluation, ranging from pleasant and positive to unpleasant and negative) and arousal (related to the feeling of energy, ranging from calm to excited). They used motion, shot length and sound energy to model the arousal dimension; valence was based solely on estimated pitch during speech segments. They indicated promising results on a single movie, but no formal evaluation was carried out.

Other authors [6] worked on detecting three different emotions (fear, sadness, joy) in movie scenes, using video features based on colour, motion and shot cut rate. Classification accuracy reached up to 80%, but only three movies were included in the test set.

A slightly larger study dealing with emotion detection in movies was reported in [12]. They mainly used colour information to detect 16 mood terms. 15 movies of different genres were included in the study, and the reported overall accuracy was also around 80%, accuracies for individual mood terms were not given.

The use of moods for movie recommendations has also gained some attention [10]. Here, the focus was on the quality of mood based recommendations, the mood tags were provided. Results improved when the moods were included in the similarity computation of movies, rather than using them as a filter criterion afterwards.

3 Data Collection

No public dataset for video mood classification exists, and most of the published work concentrates on movies rather than TV programmes. The BBC archives contain a varied mix of programmes, differing in format from quiz shows to drama and news. We therefore decided to conduct a new user study, inviting members of the general public to watch and rate video clips from the archives. After promising results from a small pilot study [4], a decision was made to conduct a large study with 200 participants.

3.1 Mood Selection

Our previous work [4] was based on a model of mood perception developed by [9]. Using a large number of semantic differentials in multiple studies, Osgood and colleagues found some reoccurring dimensions that together constituted most affective meaning. The most prominent dimension was Evaluation (measuring as how good or bad something was perceived), followed by Potency (the perception of something being strong or weak) and Activity (something being active or passive). Together these three dimensions are often referred to as EPA space.

For our study, we selected adjective pairs from Osgood's thesaurus study [9] which appeared most applicable to video. These included happy/sad and light-hearted/dark for the Evaluation dimension, serious/humorous for Potency, exciting/relaxing and fast-paced/slow-paced for Activity. We added interesting/boring, related to both Evaluation and another factor termed Receptivity. All subject ratings were given on a five point scale, with the opposing adjectives at either end.

3.2 User Trial and Video Clip Selection

The trial participants were selected to be representative of the British public in terms of age, gender, ethnicity and social background. All were watching TV at least 8 hours a week on average. The trial took place in-house at the BBC, with participants watching and rating the video clips at individual computer screens. The clips were three minutes long and preselected from a wide range of TV programmes, presented to the participants in random order.

In total, 544 programme clips were rated by at least six participants, some more clips which were only watched by a smaller number of participants were excluded from the dataset. Only one clip was extracted per programme, but a limited number of clips was taken from different episodes of the same series. Based on title matching, one clip each was extracted from 475 different series and one-off programmes, and 69 clips were extracted from multiple episodes of a further 23 series.

At the end of each session, the participants were asked if they could imagine searching for TV programmes by mood, either on its own or in combination with other search criteria. Overall, 52% of them said that they would like to be able to search by mood, 36.5% could not imagine it, and the remaining 11.5% were not sure.

4 Data Analysis

4.1 Inter-rater Agreement

As part of the basic data analysis, we evaluated the level of inter-rater agreement between subjects, using Krippendorff's Alpha [7]. The Alpha measure takes the observed user agreement and normalises it by the expected agreement, which is computed based on the relative distribution of rates. We always assumed ordinal scales, but results varied very little when the user rates were treated at interval level.

Rater agreement is shown in Fig. 1. Krippendorff's Alpha is bound between -1 and 1, a value of one means perfect agreement, zero indicates that agreement is at chance level only, and negative numbers indicate systematic disagreement. Agreement for the serious/humorous scale was highest with an alpha value of 0.69, followed by happy/sad and light-hearted/dark. Slow/fast-paced with 0.39 had a medium high rater agreement, and exciting/relaxing with only 0.23 had a relatively low agreement. It is therefore unlikely that either of these two scales corresponded directly to a single objective factor such as shot cut frequency. We had expected interesting/boring to be the mood most strongly influenced by personal preferences, this was confirmed by the low agreement of 0.20.

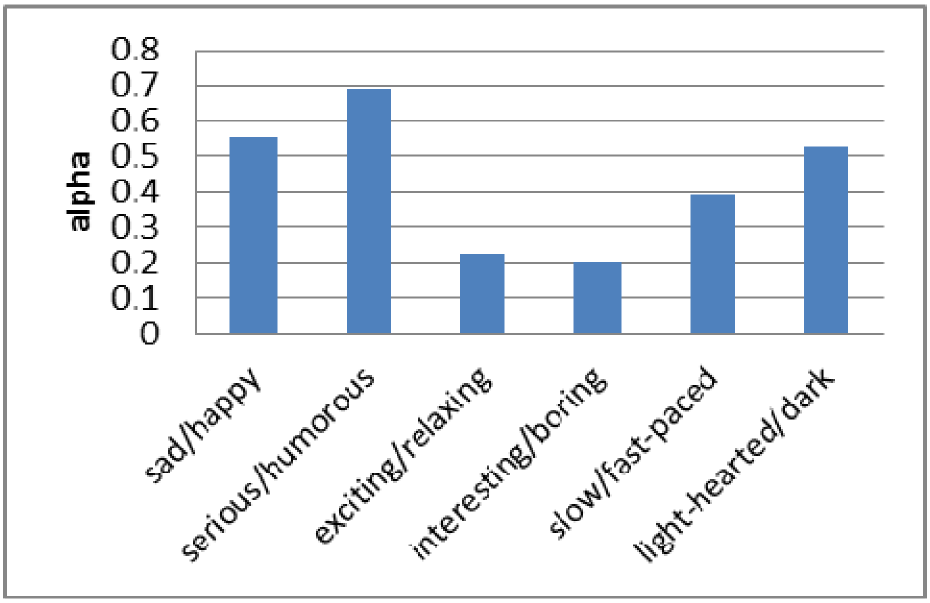

Figure 1. Rater agreement, based on Krippendorff's Alpha.

4.2 Mood Correlation

Next, we evaluated the correlation between the mood scales. With the goal of a mood based interface, highly correlated moods do not provide additional information to the user and would therefore be of little use.

All correlation values were computed using Spearman’s rank correlation [13] and are shown in Fig. 2, with adjective pairs ordered to give predominantly positive values. The highest correlation coefficient of 0.70 was found between happy/sad and humorous/serious. In the original thesaurus study by Osgood [9] these adjective pairs were clearly separated on different semantic factors, corresponding to Evaluation and Potency respectively. Previous research on mood perception of music [3] also found a high level of correlation between these two mood dimensions; the reasons for the differences to Osgood’s results remain unclear.

The correlation between the two adjective pairs chosen to represent Evaluation, happy/sad and light-hearted/dark was 0.68, also high. The correlation between the adjective pairs of the Activity dimension, exciting/relaxing and fast-paced/slow-paced was 0.44, noticeable lower, probably to some extent caused by the low rater agreement for the exciting/relaxing scale.

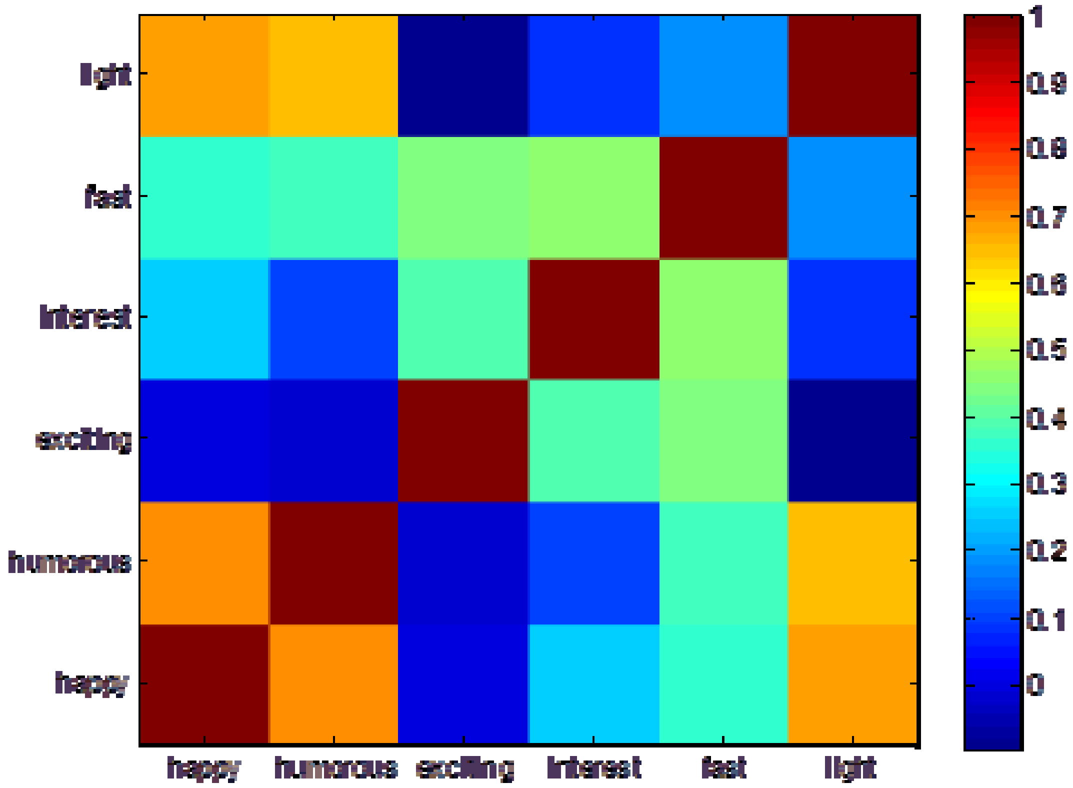

Figure 2. Correlation between moods.

4.3 Data Dimensionality

The pattern of overall correlation indicates that there might be only two independent mood dimensions in our data, rather than the three dimensions expected from the EPA model. To test this hypothesis, we conducted a principal component analysis (PCA) [13], based on averaged mood rates for each programme. Of the resulting principal components, the first contained 63% of all variance, the second 24%, and the third component less than 6%. The hypothesis of only two independent dimensions could therefore be considered correct. The first dimension of the PCA was a combination of both Evaluation and Potency, while the second axis corresponded to the Activity dimension, also including interesting/boring. A visualization of the relation of mood adjectives to PCA components is shown in Fig. 3. The location of individual programmes in the PCA space is also included, with each red dot representing a programme.

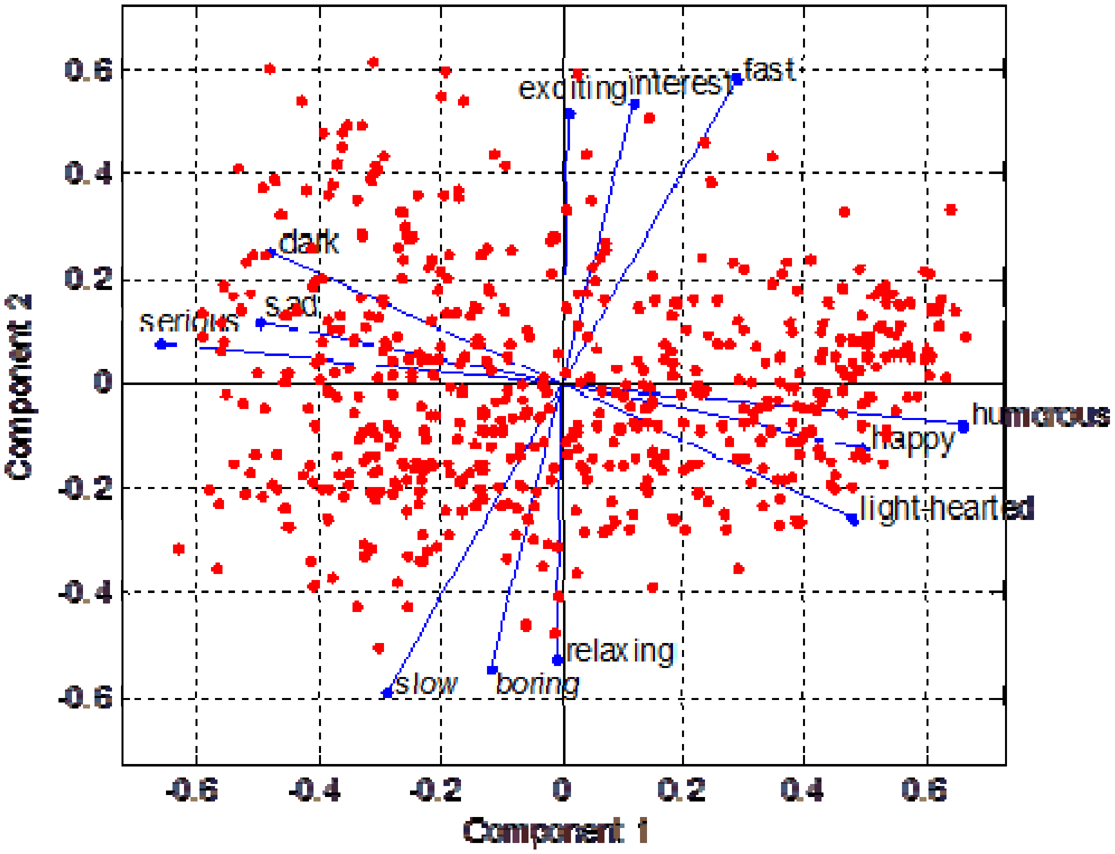

Figure 3. Moods and programmes in PCA space.

4.4 Correlation between Moods and Genres

In the special case of the BBC archives, manually assigned genre data is available for most programmes. These include formats, such as ‘Documentaries’ or ‘Competition programmes’, as well as subject genres, such as ‘Current Affairs programmes’ or ‘Comedy programmes’. The genres are hierarchically organised with up to three levels, e.g. ‘Drama programmes’ has multiple subgenres including ‘Historical drama’ and ‘Medical drama’, the latter is again subdivided into ‘Hospital drama’ and ‘Veterinary drama’. Due to inconsistencies in the human labelling, higher level categories were sometimes, but not always, assigned. In total, in our dataset we had 86 different genres, out of which 40 were subgenres of the second or third level. Each programme can have multiple genres assigned, up to four in our dataset. The assignment is binary without indication about importance or dominance of individual genres.

To evaluate the influence of genre on mood perception we computed the correlation between the genres and moods of each programme. We either used the genres directly as assigned, or exploited the hierarchical organisation of genres and also flagged all higher level genres (e.g. all ‘Hospital drama’ programmes would also be assigned the genres ‘Medical drama’ and ‘Drama programmes’).

The correlation between the moods and the binary genres using Spearman’s rank correlation [13] is shown in Fig. 4. For each mood, the genre with the highest correlation is displayed. It can be seen that the correlation is higher when higher level genres are assigned, probably caused by the general sparsity of assigned genres. Overall, the closest correlation is with comedy programmes, especially situation comedy. The serious/humorous mood scale is closely correlated with this; others like the relaxing/exciting or the boring/interesting scale have only very little correlation with the genre labels.

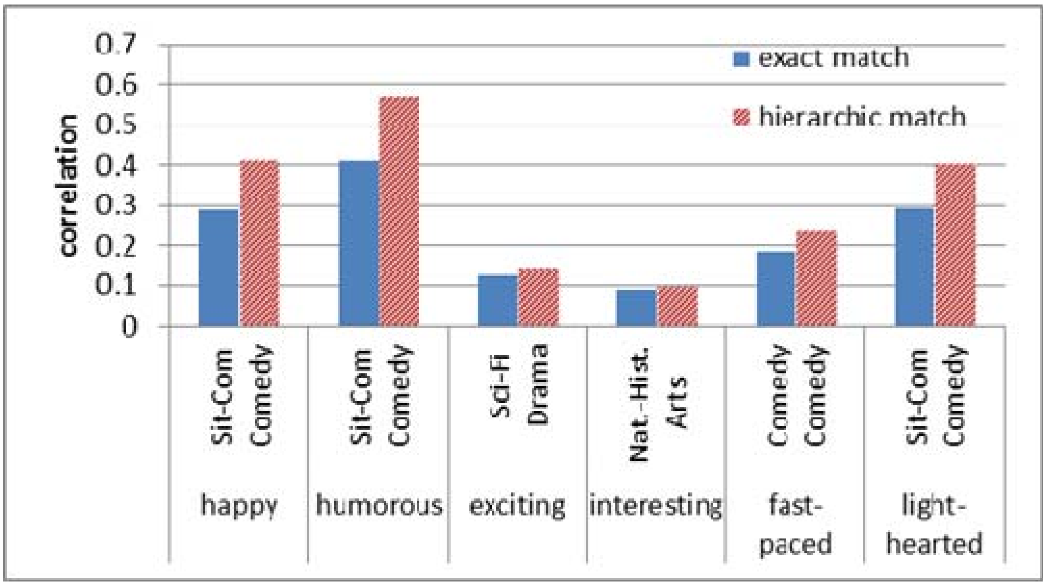

Figure 4. Correlation between moods and genre.

5 Features

5.1 Signal Processing Features

For the purpose of automatic classification, we extracted signal processing features from the video and the audio component of each programme. From the video, four different features were extracted, consisting of luminance, motion, cuts, and the constant presence of faces, details can be found in [4]. The first three features were based on downsampled versions of the video images, converted to grey scale. Luminance was computed as the average luminance of individual images. The motion feature was based on the difference between the current and the 10th preceding image, and shots boundaries were detected based on a combination of phase correlation and absolute pixel difference between the current and the previous image. The face feature was based on the face detection output from OpenCV [8]. The motivation for the face-based feature was based on our observation that the continuous presence of full frontal faces seemed to coincide with serious programmes. For each frame, the presence or absence of a face was recorded and the cumulative sum of detected faces was calculated and reset to zero whenever no face was detected. The final face feature values were scaled the by the face diameter to provide an indication of the long-term presence of full frontal faces.

Audio features were extracted using Sonic Annotator [11]. Features used were Mel-Frequency Cepstral Coefficients (MFCCs), using either C0 (overall sound energy) or the first 20 coefficients without C0. Also tested were the MFCC delta values in a range of ±10 frames, amplitude, spectral centroid, zero crossing rate, and spectral rolloff, all using a window size of 1024 samples with files having a sample rate of 48kHz.

All audio and video features were extracted on a frame by frame basis. To obtain features representing an entire video clip, the frame based features were summarised by computing the mean and standard deviation of the individual features. For each three minute clip, features consisted of mean and standard deviation of the 4 video features and 44 audio features (20 MFCCs, 20 MFCC deltas, amplitude, centroid, zero crossing, rolloff), resulting in a feature dimensionality of 96.

5.2 Genre Features

Additionally, we evaluated the existing genre information. The hierarchical labelling as described in Section IV.D slightly alleviated the sparseness problems and led to overall higher correlations with moods, and was subsequently used for all experiments. We transformed the binary genres using principal components analysis (PCA) [13]. All programmes in the training set were used to compute the PCA, and the first few components were kept as the new feature space. The number of components to use was optimised in an initial experiment and then kept for all future settings, see section VII.A.

6 Classification

6.1 Data Preparation

As a first step, we separated our data into a development and a holdout set, the latter consisting of 100 randomly chosen programmes. All experiments and parameter optimisations were based on three-fold cross-validation within the development set. The classes were evenly distributed and results varied very little between folds. Reported cross validation results are always the average across the folds. Once all parameters were fixed, we used the entire development set for training and give results on the holdout set.

For our first classification experiments, we decided to work with two moods. We chose humorous/serious and fast/slow-paced, because they corresponded to the two main components of the PCA of the mood space, showed only low correlation with each other, and had relatively high agreement between raters. We used two different settings, first we only tried to classify the extremes of each mood. We selected all programmes that had an average mood rating of two or less on the serious/humorous scale, meaning most trial participants agreed that this was clearly a serious programme. These were classified against all programmes with a rating of four or more, i.e. all clearly humorous programmes. As a result we had 185 serious and 107 humorous programmes. The same selection was independently performed for the slow-paced/fast-paced scale, for this 34 fast and 150 slow-paced examples were selected.

The second experimental setting used the average of all rates for each programme, based on the fine grained five point scale which was given to the human labellers. This means that all programmes, except for the holdout set, were included, resulting in 444 examples for both humorous/serious and fast/slow-paced. For classification results the averages were rounded to the nearest integer, resulting in five separate classes per mood. For regression, the mean values were used directly.

6.2 Machine Learning

We chose Support Vector Machines (SVMs) as our main classifier, for details and software used see [2]. SVMs were trained either for classification or regression, using radial-basis function (rbf) kernels. The main parameters (C, controlling the trade-off between model complexity and misclassification during training; ɤ, the kernel width influencing generalization abilities; and for regression ɛ, the maximum deviation allowed during training without penalisation) were optimized in a grid search. For the five class setting, we used the libSVM inbuilt extension to multi-class problems based on multiple binary classifiers.

We measured our results using the percentage of correctly classified examples. For experiments based on the five point scale, we also report root-mean-square error (rms error) [13]. We give a baseline performance, which for classification accuracy is based on choosing always the most frequent class in the training data set. For rms error, the baseline is based on always selecting the mean of all rates in the training set as predicted value for all test set items.

7 Results

7.1 Feature Selection and Mood Extremes

We started by evaluating the influence of individual features, using the simpler setup of classifying mood extremes only. All audio and video features were tested individually. When interpreting the classification results it should be noted that the class distribution was very uneven, as there were more slow than fast-paced programmes in our dataset. Picking always the most frequent class therefore already gave a high baseline accuracy of nearly 82%. For serious/humorous this was less extreme, while the dataset was slightly biased towards serious programmes, this only resulted in a baseline of 63%.

For serious/humorous, the best single audio feature were MFCCs with 90% classification accuracy. For slow/fast-paced, the energy coding coefficient C0 gave best results with 92% accuracy. Using all available audio feature improved results marginally by 1% for serious/humorous, while accuracy for slow/fast-paced actually decreased slightly by 1%. Using only video features gave lower results, the best video feature for serious/humorous was the face-based one with 69% accuracy, using all four video features increased accuracy to 79%. For slow/fast-paced, the best result of 89% accuracy was obtained using cuts, increasing slightly to 90% when all video features were included. A combination of all audio and video features gave a small increase to 93% for serious/humorous, and to 95% for slow/fast-paced. In combination with the video features, using only C0 instead of all audio features for slow/fast-cased gave slightly lower results.

We tested the genre feature on its own, varying the number of PCA components. For humorous/serious results were with 91% accuracy already very good when only the first component was used, increasing to a maximum of 94% for 32 components. For slow/fast-paced a larger number of components was needed, using only the first gave an accuracy just 1% above baseline, increasing to a maximum of 91% for 16 components.

Combining all audio, video and genre features led to 97% accuracy for serious/humorous, and to 93.5% for slow/fast-paced. For serious/humorous genre was the most important feature, achieving 94% accuracy, but similar results could also be achieved using only signal processing features, loosing just 1% accuracy.

For slow/fast-paced, signal processing features were more accurate than genre, and while genre on its own clearly contained relevant information, adding it to the signal processing features did not improve results. An overview of all results is shown in Fig. 5.

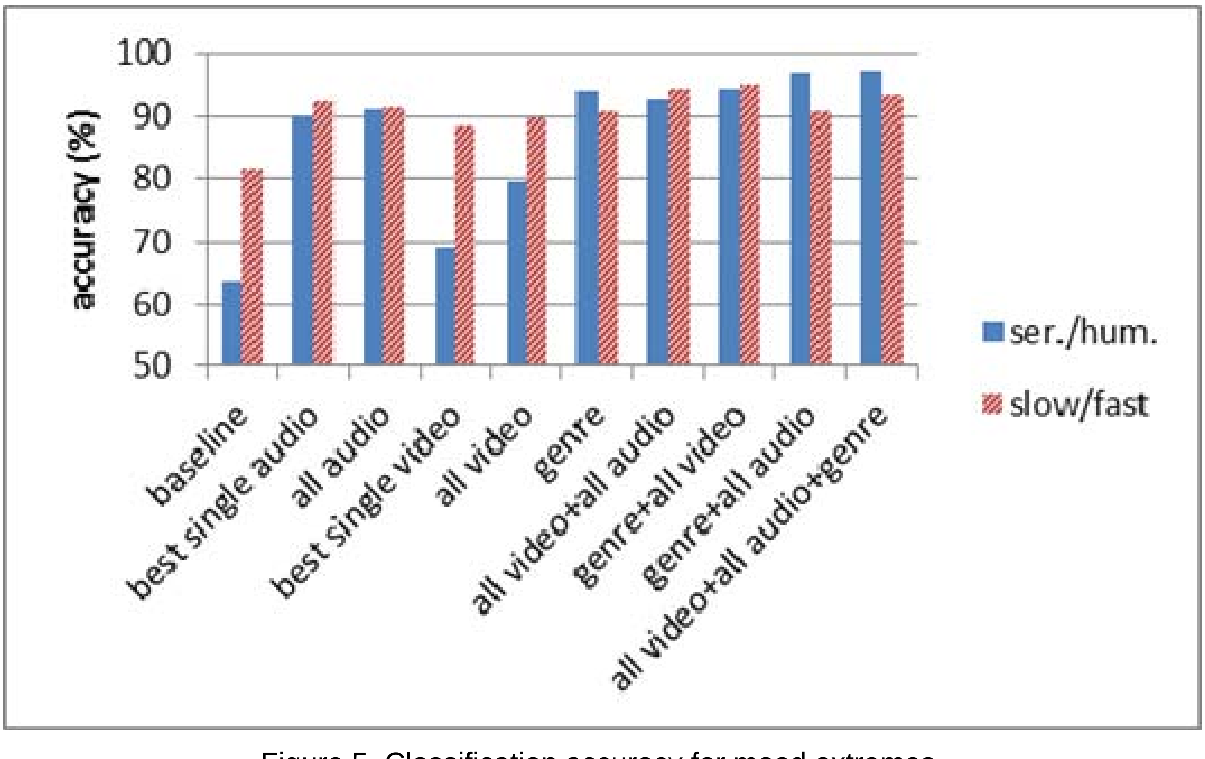

Figure 5. Classification accuracy for mood extremes.

7.2 Detailed Mood Classification

Next, we used the fine grained, averaged mood rates. Three feature set combinations were used, either entirely signal processing based using all available audio and video features, or using only genre information, or a combination of both.

In terms of classification accuracy results for the five class problem were much lower than for the extreme moods. The baseline for serious/humorous was 35%, best classification results achieved using SVMs and signal processing features were 48% accuracy, improving to 51% when genre information was added. For slow/fast-paced the baseline was 39%, best classification accuracy was 50% when signal processing features were used, increasing slightly to 51% when genre information was added.

Using SVMs trained for regression instead of classification gave similar results when the predicted regression values were rounded to the nearest class. An analysis of the rms error however showed the advantage of regression, leading to consistently lower error values than classification, see Fig. 6. The rms error for humorous/serious was 0.72 using only genre, and 0.71 when signal processing features were added, a large improvement compared to the baseline of 1.24 rms error. Using only signal processing features was less effective, leading to an rms error of 0.87. For slow/fast-paced the baseline was with 0.89 rms error much lower. Combining all features gave a slight improvement over signal processing features alone, lowering the rms error from 0.65 to 0.63.

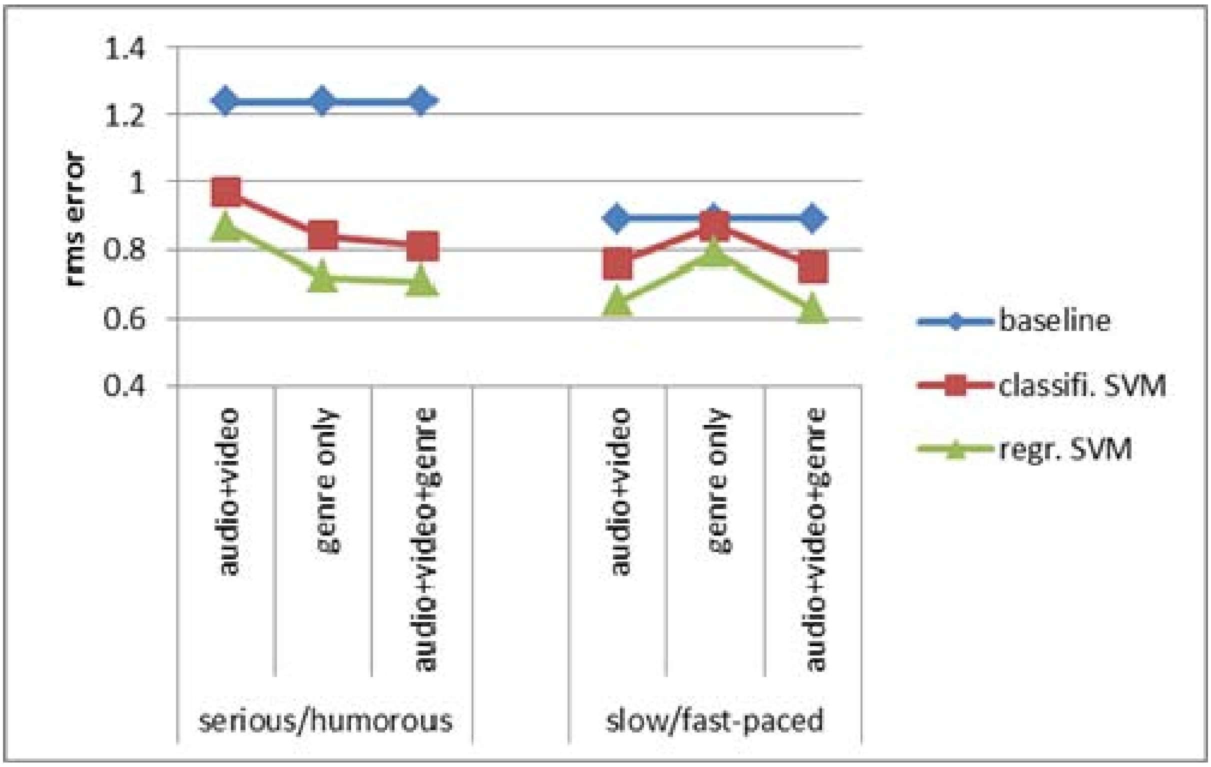

Figure 6. RMS error, mean of all rates (ranging from 1 to 5).

7.3 Holdout Set

As a last step, we evaluated results for the 100 files from the holdout set. All parameters including those for SVM training were fixed based on the best results from the cross-validation development set. We tested the holdout set using only audio and video signal processing features, or only genre information, or a combination of both. Final models were trained using the files from the development set, either those with clear moods only for the two-class setting, or all files for regression.

Within the holdout set, 43 video clips were labelled as clearly serious and 22 as clearly humorous, resulting in a baseline accuracy of 66% for the two-class setting. Using signal processing features classification accuracy was 83%, noticeable lower than the 93% achieved on the development set. Using only genre information accuracy was with 97% higher than on the development set, and a combination of genre and signal processing features was with 98% also slightly higher, but this might have been caused by the higher baseline of the holdout set.

Only 6 video clips in the holdout set were labelled as clearly fast-paced, with 30 clearly slow-paced ones, making results on this set potentially unreliable. Accuracies were with 89% accuracy for signal processing features, 86% for genre, and 92% for a combination of both well above the baseline of 83%, but nevertheless lower than for the development set, see Fig. 7.

For the regression setting based on the mean rates from all human labellers we were able use all 100 video clips from the holdout set, making the results more meaningful. Here, results for the development and the holdout set were very similar, see Fig. 8. The best rms error of for serious/humorous was 0.66 when all features were used, compared to 0.71 for the development set. Again, the better results for the holdout set can at least partly explained by the higher baseline, the improvement was with 0.53 for the development and 0.55 for the holdout set very comparable. For slow/fast-paced the rms error was nearly identical to that of the development set, giving the lowest error of 0.64 when all available features were used.

Overall, these results suggest that overfitting of parameters has not taken place and the results can be generalised.

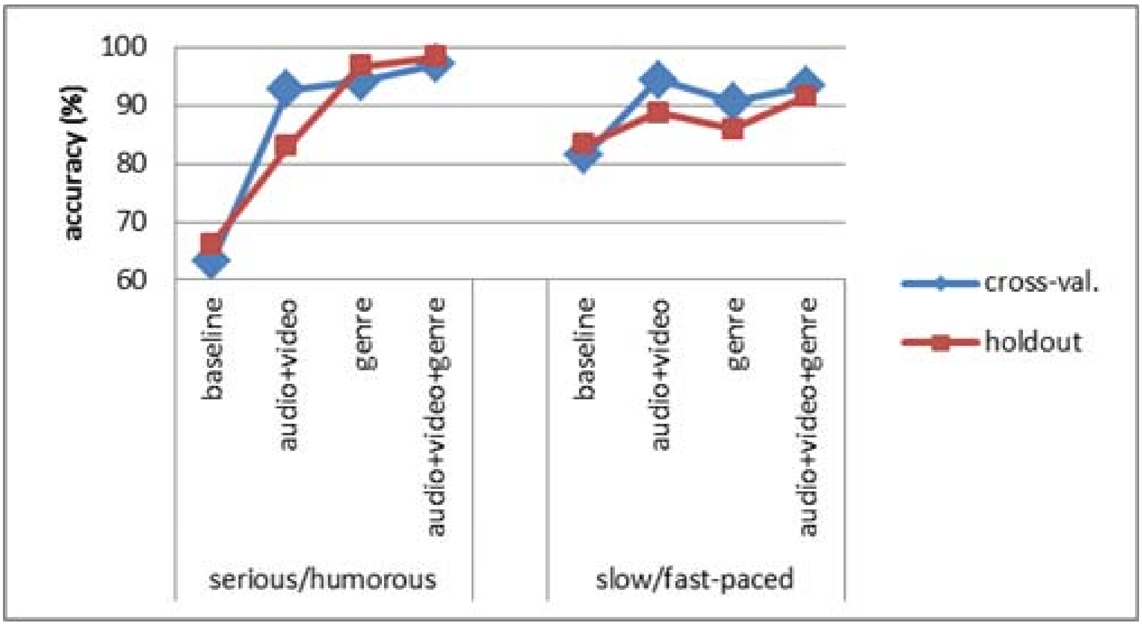

Figure 7. Classification accuracy for mood extremes,
comparing development and holdout set.

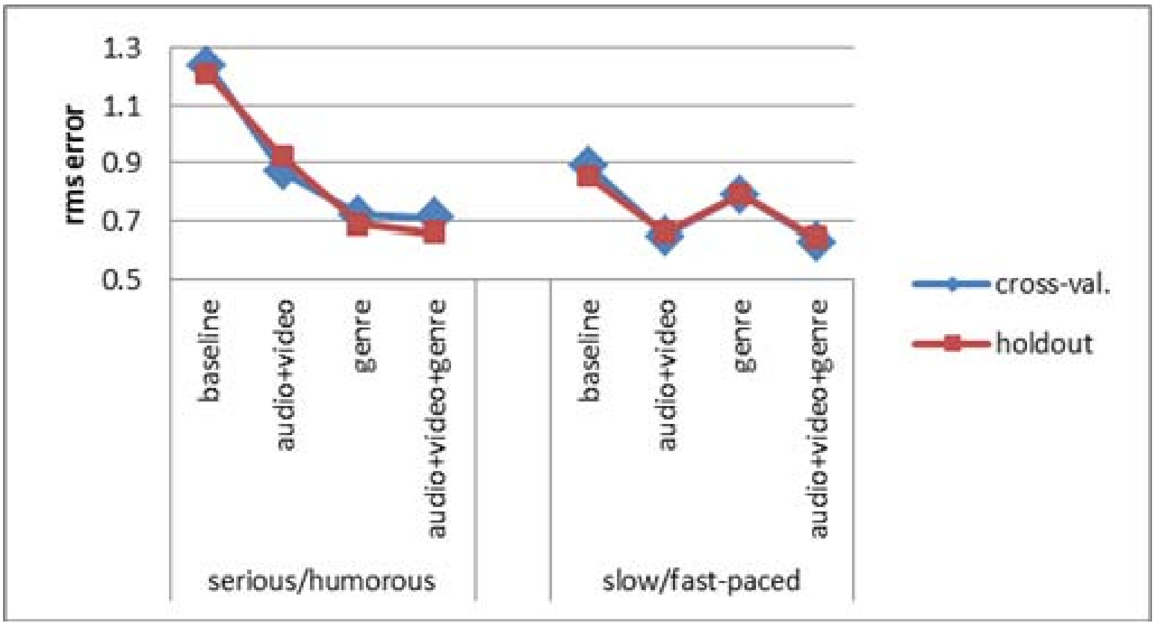

Figure 8. RMS error, mean of all rates (ranging from 1 to 5), comparing development and holdout set.

8 Conclusions and Future Work

We presented results from a large scale study of mood based classification of TV programmes, to the best of our knowledge this was the first of its kind. There was an overall agreement about which mood labels should be assigned to programmes, meaning that the perception of moods is not entirely subjective. The majority of participants said they would like to be able to search by mood. Mood perception was dominated by two independent dimensions; most important was the distinction between serious and light-hearted programmes, followed by a dimension related to perceived pace.

Automatic classification of moods was possible, even without using any human generated metadata. However, for the distinction between serious and humorous programmes, manually assigned genre information improved accuracy, and on its own was more useful than signal processing based features. The importance of human generated metadata was different for the second dimension relating to perceived pace, where genre held only limited information and signal processing features were more successful. Both dimensions could be classified with around 95% accuracy for programmes with clear moods. Regression techniques were successfully used to obtain fine graded predictions of mood values for both dimensions.

Improving classification accuracy will be part of our future work, especially for the more detailed assignment of continuous mood values. Manual inspection of the results showed that the current algorithm often failed for programmes with specific mood combinations, especially those that were both serious and fast-paced. Additional features like more accurate localised motion estimation might help to improve these results. We also want to evaluate scene based classification, as moods are likely to change over the duration of full length TV shows. A long term goal will be the implementation and evaluation of a prototype user interface for mood-based search in large archives.

9 References

[1] D. Brezeale, and D.J. Cook, “Automatic video classification: A survey of the literature,” IEEE Transactions on Systems, Man, and Cybernetics, 38 (3), 2008.

[2] C.-C. Chang, and C.-J. Lin, “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] S. Davies, P. Allen, M. Mann and T. Cox, “Musical moods: A mass participation experiment for affective classification of music”, Proc. Int. Society for Music Information Retrieval Conference, 2011.

[4] J. Eggink and D. Bland, “A pilot study for mood-based classification of TV programmes,” Proc. ACM Symposium on Applied Computing, 2012

[5] A. Hanjalic and L.-Q. Xu, “Affective video content representation and modeling," IEEE Transactions on Multimedia, 7 (1), 2005.

[6] H.-B. Kang, “Affective content detection using HMMs,“ Proc. ACM Int. Conf. on Multimedia, 2003

[7] K. Krippendorff, “Content analysis: An introduction to its methodology”, Thousand Oaks, CA: Sage, 2004.

[8] Opencv, http://opencv.willowgarage.com/wiki/FaceDetection, [Jul. 19, 2011].

[9] C.E. Osgood, G. Suci and P. Tannenbaum, “The measurement of meaning”, Uni. of Illinois Press, 1957.

[10] Y. Shi, M. Larson, A. Hanjalic, “Mining mood-specific movie similarity with matrix factorization for context-aware recommendation,” Proc. Challenge on Context-aware Movie Recommendation, 2010

[11] Sonic annotator, http://www.omras2.org/SonicAnnotator, [May 04, 2011]

[12] C.-Y. Wei, N. Dimitrova, and S.-F. Chang, “Color-mood analysis of films based on syntactic and psychological models,” Proc. IEEE Int. Conf. on Multimedia and Expo, 2004.

[13] Wikipedia, http://en.wikipedia.org/wiki/Principal_component_analysis, /Root_mean_square_error, /Spearman_rank_correlation [Nov. 16, 2011]