A KNOWLEDGE TRANSFER AND BOOSTING APPROACH TO THE PREDICTION OF AFFECT IN MOVIES

Sabyasachee Baruah[o], Rahul Gupta[+], Shrikanth Narayanan[+]

[o]Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India
[+]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA

ABSTRACT

Affect prediction is a classical problem and has recently garnered special interest in multimedia applications. Affect prediction in movies is one such domain, potentially aiding the design as well as the impact analysis of movies. Given the large diversity in movies (such as different genres and languages), obtaining a comprehensive movie dataset for modeling affect is challenging while models trained on smaller datasets may not generalize. In this paper, we address the problem of continuous affect ratings with the availability of limited in-domain data resources. We initially setup several baseline models trained on in-domain data, followed by a proposal of a Knowledge Transfer (KT) + Gradient Boosting (GB) approach. KT learns models on a larger (mismatched) data which are then adapted to make predictions on the data of interest. GB further updates these predictions based on models learnt from the in-domain data. We observe that the KT + GB models provide Concordance Correlation Coefficient values of 0.13 and 0.27 for valence and affect prediction on the *continuous LIRIS ACCEDE* dataset against best baseline prediction values of 0.12 and 0.11. Not only the KT + GB models improve the overall performance metrics, we also observe a more consistent model performance across movies of various genres.

Index Terms— Gradient Boosting, Knowledge Transfer, affect prediction in movies

1. INTRODUCTION

Movies are often (if not always) designed to convey a specific emotional experience on its audience [1, 2]. Researchers have extended the classical problem of affect prediction to the domain of movies with the goal of understanding the impact of movies as well as aiding the design and analysis of movies [3, 4]. Most of the existing algorithms make use of statistical models trained on audio-visual features to predict affective dimensions (e.g. valence and arousal). These statistical models often require sufficiently large amounts of data for a low generalization error [5]. Given that movies span a vast variety of genres, are recorded in different languages and differ across cultures, comprehensive movie datasets (along with the desired set of annotations) may not always be available to train such statistical models. In this paper, we address the problem of continuous affect tracking in movies, with limited availability of in-domain data to train low error statistical models. We propose a Knowledge Transfer (KT) approach aided with Gradient Boosting (GB) to predict affective dimensions in the *continuous LIRIS-ACCEDE (Annotated Creative Commons Emotional Database for affective video content analysis)* database [6]. We train the KT models on a larger (albeit mismatched) dataset, which are later adapted to the dataset of interest. GB models are trained on the smaller in-domain data and operate along with KT models to predict affect induced by movies (thereby accumulating information learnt on in-domain as well as out-of-domain resources). The overarching goal of our experiments is to unify the ongoing efforts in multimedia related research, potentially sharing resources despite inherent incompatibilities.

Several previous works have investigated the emotional impact of movies [7, 8]. Furthermore, researchers have also applied machine-learning algorithms to both understand and predict emotions induced by movies. Examples include movie content analysis based on arousal and valence features [9], affect ranking of movie scenes using physiological signals and content analysis [10] as well as examination of other supervised methods [11], deep learning and kernel methods [12] within the prediction of affect in movies. Similarly, studies have also addressed the issue of robustness in affect prediction using mixture models [13] and multimodal learning [14]. On the other hand, studies have investigated knowledge transfer (/transfer learning) and have proposed various algorithms such as boosting [15], and transfer learning via dimensionality reduction [16]. KT methods have been applied to several domain such as text classification [17] and cross-language classification [18]. In this work, we propose a knowledge transfer approach, aided with gradient boosting, to predict affective dimensions in movies. This approach combines learning from external as well as in-domain resources. To the best of our knowledge, this is the first such work in movie affect prediction (particularly using data with mismatch in content as well as the annotation protocol and granularity).

The *continuous LIRIS-ACCEDE* dataset consists of 30 movies spanning various genres, languages as well as composition (e.g. color vs gray-scale and live-action vs animated). We initially train various baseline statistical models on the limited in-domain dataset to predict the affective dimensions of valence and arousal at a frame rate of 1 value per second. We analyze the baseline results, observe the performance of models across the movies and motivate the application of KT + GB models. The KT models are trained on a larger dataset and the model predictions are then adapted to perform prediction on the dataset of interest. We then propose a GB approach, incorporating KT models and models trained on in-domain data as components. The GB models are trained sequentially on a pseudo-residual from the previous models and usually are ensemble of weak learners [19]. Our results reflect that KT models, on their own, improve upon the conventional in-domain statistical models. While the performance for valence saturates using KT models, GB models provide further leverage using in-domain data for arousal. The proposed KT + GB models achieve Concordance Correlation Coefficient (CCC, a metric accounting for both correlation and bias difference [20, 21]) values of 0.13 and 0.27 between true and predicted valence and arousal ratings, respectively (against best baseline performances of 0.12 and 0.11). Further analysis also reveals that not only KT + GB models enhance the performance, but their performances are more consistent across individual movies in the dataset. We describe our dataset of choice in the next section, followed by a description of our methodology.

2. DATASET

We use the *continuous Annotated Creative Commons Emotional Database for affective video content analysis* (*continuous LIRIS-*

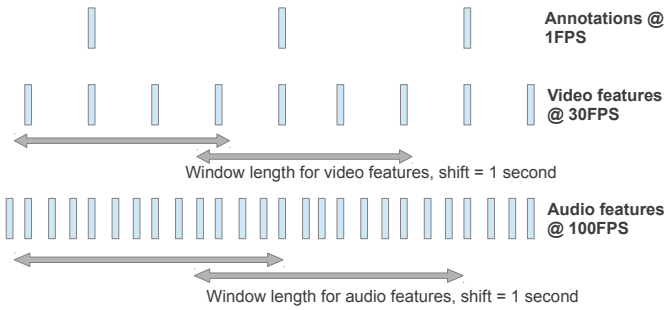

Fig. 1. Feature extraction scheme using extraction of statistics over a temporal window of audio/video frames. Length of temporal window is the one that maximizes mutual information of features with the annotations.

Table 1. List of features extracted on the *continuous LIRIS-ACCEDE* dataset

Source	Features	FPS	Toolbox
Visual	Luminance, intensity and optical flow	30	OpenCV [24]
Audio	Mel-frequency cepstral coefficients, coefficients, voicing probability, harmonic to noise ratio, zero crossing rate, crossing rate, fundamental frequency, log energy	100	OpenSmile [25]
Music	Chroma features (12 semitones)		

ACCEDE) dataset for the purpose of our experiments. This dataset was also used as part of the *emotional impact of movies task* organized at the MediaEval challenge 2015 [22]. The dataset consists of 30 short films of length varying from 3 to 28 minutes. A set of ten annotators rate each movie for the affective dimensions of arousal and valence at a frame rate of 1 sample per second, within a range of -1 to 1. The final ratings are obtained as frame-wise mean of annotations from each annotator. For further information regarding the dataset, we refer the reader to [6]. In the next section, we describe the features used in our experiments.

2.1. Feature extraction

We use an assembly of visual, speech and music features to train our models. Table 1 shows the list of features along with their frame rate and the toolbox used. Note that the frame rates of features from different sources are different. In order to synchronize the features with the annotations, we compute statistics on the features along a temporal window with a shift of 1 second (thereby obtaining a feature vector per second). We compute a set of nine statistics (mean, median, standard deviation, kurtosis, lower and upper quartile, minimum, maximum and range) for every feature. Figure 1 depicts the extraction of these statistics on the features. We choose the length of the temporal windows by maximizing the mutual information between the computed features and affective dimension labels (available on the training set in a cross-validation framework described in the next section). The mutual information computation assumes a Gaussian distribution for annotations and features, and is implemented as suggested in [23] (Section 4). We denote the set of features for a given movie $\mathcal{F}$ as the vector $\mathbf{X}_{\mathcal{F}} = [X_{\mathcal{F}}(1), .., X_{\mathcal{F}}(n), .., X_{\mathcal{F}}(N)]$, where N is the total number of analysis frames (equals annotation frames) and $X_{\mathcal{F}}$(n) is the set of audio and video features extracted for the analysis frame n. The ground truth annotations are represented as $\mathbf{t}_{\mathcal{F}} = [t_{\mathcal{F}}(1), .., t_{\mathcal{F}}(n), .., t_{\mathcal{F}}(N)]$, where $t_{\mathcal{F}}(n)$ is the ground truth annotation for the n^{th} frame.

3. METHODOLOGY

We test several regression schemes to predict the continuous valence and affect ratings from the features described in the previous section. Initially, we establish a baseline using multiple regression schemes. We discuss the performance of the baseline models and motivate our KT and GB based method. All our experiments are performed by using a leave one movie out cross-validation scheme. During cross-validation, we use 25 movies as the training set, 4 for validation and 1 movie in the test set. The primary metric for the evaluation of our methods is Concordance Correlation Coefficient (CCC) [20] between the annotated ground truth and the predicted ratings on the test set. This metric accounts for both the similarity of pattern between the two time series as well as the bias difference between them. CCC has been used as a metric in several recent time series prediction experiments [21, 26]. Next, we describe the baseline models.

3.1. Baseline

We use a set of three regressors (linear regression, ridge regression and neural network) to predict the affect ratings from the features. Linear regression [27] is the simplest of the three regressors and linearly maps the features to the affect ratings. Like linear regression, ridge regression [27] also performs a linear mapping. However the weights for ridge regression are regularized [27], a scheme helpful in cases involving limited amount of training data. We test two version of these schemes: with and without feature selection. During feature selection, we remove features with absolute value of the correlation coefficient below a certain threshold (tuned on the development set). Correlation coefficient quantifies the linear relationship between the features and the ratings, therefore removal of features with a low correlation coefficient can potentially boost the performance of linear models such as linear and ridge regression. Finally, neural networks [27] perform a non-linear mapping between the features and the affect ratings. We train a neural network with one hidden layer. The number of neurons is tuned on the development set and they have a sigmoidal activation. We do not perform the correlation coefficient based feature selection with neural networks as it can model non-linear relations between the features and the final labels. Table 3 shows the results for the performance of these three baseline regression models (the table is shown in results section (Section 4) for ease of comparison with other proposed methods).

Discussion: From the results, we observe that the performance for affect prediction varies across the three models. Neural network regressor performs the best for valence, while linear regression with feature selection performs the best for arousal. This indicates that a non-linear mapping performs better for valence prediction, while arousal is predicted best by a simpler linear model. We further list the performance for each movie as obtained using the best regressors for both arousal and valence in Figure 3. The figure shows that the performance per movie varies quite a bit, indicating high error variation depending upon the movie at hand. For further analysis, we show the histogram of movie genre distribution in the *continuous LIRIS-ACCEDE* dataset in Figure 2. We observe that certain genres are poorly represented in the dataset, for instance, romance. This is also reflected in our results as the trained models do not perform well on the only romantic movie in the database: "To Claire from Sonny" (marked in Figure 3) as the CCC values for both arousal and valence predictions are negative for this movie. This analysis reflects that a limited data representation despite large diversity in our dataset affects the robustness of the baseline models. We propose a knowledge transfer + gradient boosting methodology to address this problem, as is discussed in the next section.

3.2. Proposed method

As observed from the results and analysis in the previous section, we need to address the problem of limited data leading to poor gen-

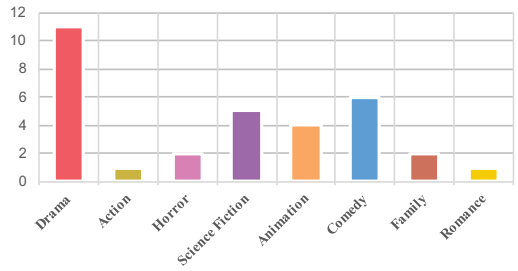

Fig. 2. Distribution of genre in *continuous LIRIS-ACCEDE* dataset

Table 2. Differences between *continuous LIRIS-ACCEDE* and *discrete LIRIS-ACCEDE*.

	Continuous LIRIS-ACCEDE	Discrete LIRIS-ACCEDE
Duration	3 - 28 minutes	8 - 12 seconds
Annotation granularity	Continuous (1 value/second)	Global
Range	-1 to 1	1 to 5

eralization of our models. In this section, we propose a knowledge transfer combined with a gradient boosting approach to address this issue. During KT, we borrow information learnt on a larger (albeit mismatched) dataset for affect prediction on the dataset of interest. This is followed by GB, where we combine models learnt on the *continuous LIRIS-ACCEDE* with the existing KT models. We discuss these proposed models in detail below.

3.3. Knowledge Transfer (KT)

During KT, we train models on a another dataset and operate them on the *continuous LIRIS-ACCEDE* dataset to obtain the affect ratings. To this effect, we use the *discrete LIRIS-ACCEDE* dataset consisting of a larger set of ∼10k movie clips. The *discrete LIRIS-ACCEDE* dataset is different from the *continuous LIRIS-ACCEDE* dataset in several respects, as summarized in Table 2. The discrete dataset consists of small movie clips (∼10 seconds in length) as opposed to the continuous dataset (3-28 mins in length). Furthermore, the annotations are provided at the global scale with a single valence and arousal rating for each clip. The annotations are in the range of 1-5 as opposed to the scale of -1 to 1 in the continuous dataset. We describe the model training on the discrete dataset and application on continuous dataset in detail below.

3.3.1. KT model training

Initially, we extract the same of set of features on the discrete dataset as mentioned in the Table 1. However, we compute the statistics (listed in section 2.1) over the entire duration of the clip (unlike the window-wise approach for *continuous LIRIS-ACCEDE* dataset). We then train a regression model on the feature statistics to predict the discrete ratings. The choice for regression model is empirically determined to be a ridge regressor based on a 3-fold cross-validation on the *discrete LIRIS-ACCEDE* dataset. Its parameters are also tuned using the 3-fold cross-validation.

3.3.2. KT model application

In order to apply the model trained on the discrete dataset, we re-use the feature assembly in Table 1. However the feature statistics are computed over a fixed window length of 10 seconds with a shift of 1 second (see Figure 1). This window length is empirically chosen to match the length of clips in the discrete dataset. We then predict the affective dimensions per second using the model trained in the last step. One can view this operation as predicting global affect at a sliding window level from the continuous data using the KT model. Since the annotation scales are different for each dataset, we linearly scale the ratings predicted by the KT model. We obtain the predictions using the KT model on the training set and learn a linear scaling such that the minimum mean squared error between the KT model predictions and the true ratings is minimized. This scaling is then applied on the testing set during evaluation. After the KT step, we train a gradient boosting model as discussed in the next section.

3.4. Gradient boosting (GB)

The goal of GB model is to obtain information from the smaller set of *continuous LIRIS-ACCEDE* dataset and incorporate it along with the outputs from the KT model for final affect prediction. We use the gradient boosting approach similar to the one proposed by Gupta et al. [28] for our experiments. Gupta et al. [28] proposed an approach using linear filters as base learners to predict continuous affect in music. We use a modified approach, where we minimize mean squared error using linear regression on the feature statistics after introducing a temporal delay. We refer the reader to [19, 28] for the details of gradient boosting in minimizing the mean squared error and briefly describe the GB model used in our experiments below.

3.4.1. Gradient boosting algorithm for affect prediction

The proposed gradient boosting learns an ensemble of $K+1$ base learners $\{\tilde{\boldsymbol{h}}_0, \tilde{\boldsymbol{h}}_1, .., \tilde{\boldsymbol{h}}_K\}$, represented as $\boldsymbol{M}_K$. For the set of features $\boldsymbol{X}_\mathcal{F}$ for movie $\mathcal{F}$, the affective predictions $\boldsymbol{M}_K(X_\mathcal{F})$ are computed as

$$\boldsymbol{M}_K(\boldsymbol{X}_\mathcal{F}) = \sum_{k=0}^{K} \tilde{\boldsymbol{h}}_k(\boldsymbol{X}_\mathcal{F}) \tag{1}$$

where $\tilde{\boldsymbol{h}}_k(\boldsymbol{X}_\mathcal{F})$ is the prediction from the k^{th} base learner. We represent the mean squared error for the ensemble of learners $\boldsymbol{M}_K(\boldsymbol{X}_\mathcal{F})$ as $\boldsymbol{E}$, computed as shown below:

$$\boldsymbol{E} = \sum_{\mathcal{F} \in \text{ training set}} \frac{1}{2} ||\boldsymbol{t}_\mathcal{F} - \boldsymbol{M}_K(\boldsymbol{X}_\mathcal{F})||_2^2 \tag{2}$$

Each of the base learners $\{\tilde{\boldsymbol{h}}_0, \tilde{\boldsymbol{h}}_1, .., \tilde{\boldsymbol{h}}_K\}$ is learnt using the following algorithm.

• The base model $\boldsymbol{M}_0 = \tilde{\boldsymbol{h}}_0$ is set as the KT model. Hence the predictions $\boldsymbol{M}_0(\boldsymbol{X}_\mathcal{F})$ are obtained from the KT model described in the previous section. The subsequent predictions are added to $\tilde{\boldsymbol{h}}_0(\boldsymbol{X}_\mathcal{F})$ in a boosting fashion, therefore integrating the knowledge learnt from KT and current data.

• For k = 1 to K

– Compute the pseudo-residuals $\boldsymbol{r}_\mathcal{F}^k = [r_\mathcal{F}^k(1), .., r_\mathcal{F}^k(n), .., r_\mathcal{F}^k(N)]$ for each movie $\mathcal{F}$ in the training set, where

$$\boldsymbol{r}_\mathcal{F}^k = -\frac{\partial \boldsymbol{E}}{\partial \boldsymbol{M}_k(\boldsymbol{X}_\mathcal{F})} = -\frac{\partial \left(\frac{1}{2} ||\boldsymbol{t}_\mathcal{F} - \boldsymbol{M}_k(\boldsymbol{X}_\mathcal{F})||_2^2 \right)}{\partial \boldsymbol{M}_k(\boldsymbol{X}_\mathcal{F})} \Big|_{\text{at } \boldsymbol{M}(\boldsymbol{X}_\mathcal{F}) = \boldsymbol{M}_k(\boldsymbol{X}_\mathcal{F})}$$
$$= \boldsymbol{t}_\mathcal{F} - \boldsymbol{M}_k(\boldsymbol{X}_\mathcal{F}) \tag{3}$$

– Compute a temporal shift in the features that maximizes the mutual information between the shifted features and the pseudo-residuals $\boldsymbol{r}_\mathcal{F}^k$, computed across all movies in the training set. The shift is computed using Gaussian assumption and methodology suggested in [23] (Section 4).

– Train a linear regressor $\boldsymbol{h}_k$ to predict the pseudo-residual using shifted features. The motivation of using shifted features in training $\boldsymbol{h}_k$ is to allow for affect prediction using feature values from multiple time stamps. During testing, same shift is applied to the testing set features before applying $\boldsymbol{h}_k$.

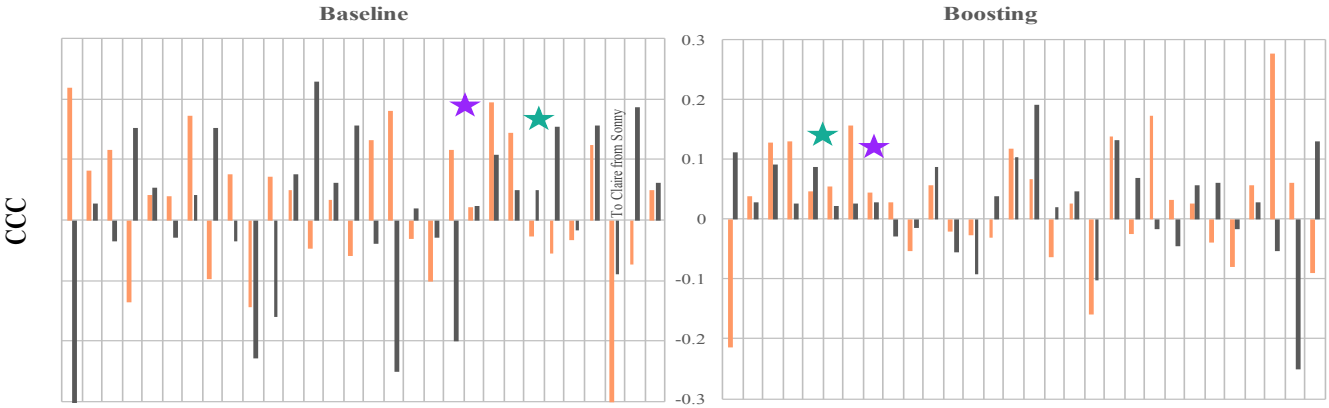

Fig. 3. Performance of the best baseline model (left) and KT + GB method (right) across the 30 movies in the *continuous LIRIS-ACCEDE* dataset. A more variation in performance is observed for the baseline models. The green and purple stars denote the standard deviation for valence and arousal CCC values, respectively .

– Compute weight γ_k for the linear regressor learnt in the last step using the following one-dimensional optimization problem. γ_k is the weight used to scale the outputs from the regressor.

$$\gamma_k = \arg\min_{\gamma} \sum_{\substack{\mathcal{F}\in \\ \text{training set}}} ||\boldsymbol{t}_{\mathcal{F}} - \Big(\boldsymbol{M}_{k-1}(\boldsymbol{X}_{\mathcal{F}}) + \gamma_k \times \boldsymbol{h}_k(\boldsymbol{X}_{\mathcal{F}})\Big)||_2^2 \tag{4}$$

– Update the model.

$$\boldsymbol{M}_k(\boldsymbol{X}_{\mathcal{F}}) = \boldsymbol{M}_{k-1}(\boldsymbol{X}_{\mathcal{F}}) + \tilde{\boldsymbol{h}}_k(\boldsymbol{X}_{\mathcal{F}}) = \boldsymbol{M}_{k-1}(\boldsymbol{X}_{\mathcal{F}}) + \big(\gamma_k \times \boldsymbol{h}_k(\boldsymbol{X}_{\mathcal{F}})\big) \tag{5}$$

The number of base learners K is tuned on the development set. We summarize the results of the KT and GB models and discuss them in the next section. The GB models incorporates KT model and is referred to as KT + GB model.

4. RESULTS AND DISCUSSION

Table 4 shows the results for KT and GB models, along with the baseline results in Table 3. From the results, we observe that the KT model on its own outperforms the baseline. This indicates that affect modeling from a larger (and mismatched) dataset outperforms the models learnt on a smaller in-domain dataset. Despite being mismatched, the larger dataset contains more representative data, thereby improving performance on the data of interest. Training the model further using GB incorporating KT model as the first base learner improves performance for arousal. However, an improvement is not observed for valence prediction. This may indicate that valence prediction based on the limited amount of in-domain data does not improve the performance beyond that predicted by the KT model. We perform further analysis regarding the performance of the models on each movie individually, and present our findings.

Discussion: In order to compare with the per-movie performance of the baseline system, we present the performance of the KT + GB model in Figure 3 for each movie. From the figure, we observe that the performance of the movies is more consistent in the KT + GB models, with fewer movies showing a negative CCC in arousal and valence. Further, we compare the standard deviations of per-movie arousal (σ_{aro}) and valence (σ_{val}) performances between the baseline and KT + GB models (σ computed over arousal and valence CCC for the 30 movies). σ_{aro} computed over KT + GB model CCC values is significantly lower than when computed over best baseline model CCC values (F-test, p-value$<$5%). Although, the reduction in (σ_{val}) using GB models over the best baseline model is not significant (F-test, p-value = 0.18), we do observe a marginal decrease.

Table 3. CCC values for affect prediction using the baseline regressors. The best performances for each dimension are shown in bold. FS indicates training with feature selection.

	Valence	Arousal
Linear regression	0.06	0.06
Linear regression + FS	0.07	**0.11**
Ridge regression	0.04	0.03
Ridge regression + FS	0.04	0.10
Neural networks	**0.12**	0.02

Table 4. CCC values for affect prediction using the KT + GB.

	Valence	Arousal
Knowledge transfer	0.13	0.22
Gradient boosting	0.13	0.27

Overall, our experiments suggest that using KT + GB models outperform the baseline models and the motivation for their use lies in the limited availability of in-domain data. Our models not only improve the performance of affect prediction, but are also more consistent in predicting affect across the movies in the dataset, sampled from various genres. In the next section, we present our conclusions and a few future directions.

5. CONCLUSION

Research has extended affect prediction to assess the emotional impact of movies, potentially aiding design and analysis of movies. However, statistical models often require large amounts of data to achieve a low error performance. This issue is further complicated by the vast diversity in movies. We propose a KT + GB approach in this paper to address this issue. KT models borrow knowledge from other larger dataset and GB models incorporate KT models along with models learnt on the in-domain data. We not only demonstrate the superior performance of KT + GB models on the *continuous LIRIS-ACCEDE* dataset, but also achieve a more consistent performance across movies.

In the future, we aim to extend the current work to other time series data for media analysis such as interestingness [29] and event prediction [30]. From the point of view of modeling, we aim to explore other options such as incorporating CCC as a direct optimization cost (current GB models use squared error as a proxy), exploring other modeling schemes (e.g. neural networks) within the KT + GB models. Finally, we also aim to extend the models to other domains with similar issues such as affect in music [31] and engagement prediction [32].

6. REFERENCES

[1] A. Bartsch, "Emotional gratification in entertainment experience. why viewers of movies and television series find it rewarding to experience emotions," *Media Psychology*, vol. 15, no. 3, pp. 267–302, 2012.

[2] L. Canini, S. Benini, P. Migliorati, and R. Leonardi, "Emotional identity of movies," in *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2009, pp. 1821–1824.

[3] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A dimensional approach to emotion recognition of speech from movies," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 65–68.

[4] Y.-G. Jiang, B. Xu, and X. Xue, "Predicting emotions in user-generated videos." in *AAAI*, 2014, pp. 73–79.

[5] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1.

[6] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[7] A. Bartsch, M. Appel, and D. Storch, "Predicting emotions and meta-emotions at the movies: The role of the need for affect in audiences experience of horror and drama," *Communication Research*, vol. 37, no. 2, pp. 167–190, 2010.

[8] N. Carroll, "Movies, the moral emotions, and sympathy," *Midwest Studies in Philosophy*, vol. 34, no. 1, pp. 1–19, 2010.

[9] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 677–680.

[10] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *Proceedings of the 2nd ACM workshop on Multimedia semantics*. ACM, 2008, pp. 32–39.

[11] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 2376–2379.

[12] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "Deep learning vs. kernel methods: Performance for emotion prediction in videos," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 77–83.

[13] A. Goyal, N. Kumar, T. Guha, and S. Narayanan, "A multimodal mixture-of-experts model for dynamic emotion prediction in movies," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2016.

[14] L. Pang and C.-W. Ngo, "Mutlimodal learning with deep boltzmann machine for emotion prediction in user generated videos," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 619–622.

[15] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 193–200.

[16] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction." in *AAAI*, vol. 8, 2008, pp. 677–682.

[17] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, "Transferring naive bayes classifiers for text classification," in *Proceedings of the national conference on artificial intelligence*, vol. 22, no. 1. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007, p. 540.

[18] X. Ling, G.-R. Xue, W. Dai, Y. Jiang, Q. Yang, and Y. Yu, "Can chinese web pages be classified with english data source?" in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 969–978.

[19] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[20] J. J. Liao and J. W. Lewis, "A note on concordance correlation coefficient," *PDA Journal of Pharmaceutical Science and Technology*, vol. 54, no. 1, pp. 23–26, 2000.

[21] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016-depression, mood, and emotion recognition workshop and challenge," *arXiv preprint arXiv:1605.01600*, 2016.

[22] M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen, "The mediaeval 2015 affective impact of movies task," in *MediaEval 2015 Workshop*, 2015.

[23] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, 2015.

[24] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[26] S. Krishna, R. Gupta, M. Nasir, B. Booth, S. Lee, and S. Narayanan, "Online affect tracking with multimodal kalman filters," in *6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016.

[27] C. M. Bishop, "Pattern recognition," *Machine Learning*, vol. 128, 2006.

[28] R. Gupta, N. Kumar, and S. Narayanan, "Affect prediction in music using boosted ensemble of filters," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 11–15.

[29] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su, "Violence detection in movies," in *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on*. IEEE, 2011, pp. 119–124.

[30] D. S. Butterfield, C. Fake, C. J. Henderson-Begg, and S. Mourachov, "Interestingness ranking of media objects," Aug. 23 2012, uS Patent App. 13/593,112.

[31] J. Broekens, A. Pronker, and M. Neuteboom, "Real time labeling of affect in music using the affectbutton," in *Proceedings of the 3rd international workshop on Affective interaction in natural environments*. ACM, 2010, pp. 21–26.

[32] R. Gupta, D. Bone, S. Lee, and S. Narayanan, "Analysis of engagement behavior in children during dyadic interactions using prosodic cues," *Computer Speech & Language*, vol. 37, pp. 47–66, 2016.