A Multi-task Learning Framework for Time-continuous Emotion Estimation from Crowd Annotations

Mojtaba Khomami Abadi[1,2], Azad Abad[1], Ramanathan Subramanian[3], Negar Rostamzadeh[1], Elisa Ricci[4,5], Jagannadan Varadarajan[3], Nicu Sebe[1]
[1]Department of Information Engineering and Computer Science, University of Trento, Italy.
[2]Semantic, Knowledge and Innovation Lab (SKIL), Telecom Italia, Trento, Italy.
[3]Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore.
[4]University of Perugia, Italy [5]Fondazione Bruno Kessler, Trento, Italy.
khomamiabadi,abad,rostamzadeh,sebe@disi.unitn.it, eliricci@fbk.eu, Subramanian.R,vjagan@adsc.com.sg

ABSTRACT

We propose Multi-task learning (MTL) for time-continuous or dynamic emotion (valence and arousal) estimation in movie scenes. Since compiling annotated training data for dynamic emotion prediction is tedious, we employ crowdsourcing for the same. Even though the crowdworkers come from various demographics, we demonstrate that MTL can effectively discover (1) consistent patterns in their dynamic emotion perception, and (2) the low-level audio and video features that contribute to their valence, arousal (VA) elicitation. Finally, we show that MTL-based regression models, which simultaneously learn the relationship between low-level audio-visual features and high-level VA ratings from a collection of movie scenes, can predict VA ratings for time-contiguous snippets from each scene more effectively than scene-specific models.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing; I.5.2 [**Pattern Recognition Design Methodology**]: Pattern analysis

General Terms

Measurement, Algorithms, Verification, Human Factors

Keywords

Multi-task Learning, Time-continuous emotion estimation, Crowd annotation, Movie clips

1. INTRODUCTION

Affective video tagging has been acknowledged as an important multimedia problem for long, given its utility for applications such as personalized media recommendation. However, most content and user-based media tagging approaches seek to recognize the *general* emotion of a stimulus (typically a movie or audio/video music clip), and only a few methods such as [7] have attempted to determine the dynamic of time-continuous emotion profile in the stimulus. This limitation is partly attributed to the fact that *interpreting* and *measuring* emotion is an inherently difficult problem– emotion is a highly subjective feeling, and the discrepancy between the emotion *envisioned* by the content creator versus the actual emotion *evoked* in consumers has been highlighted by many works. Also, learning the relationship between low-level content (typically in the form of audio-visual effects) and the high-level emotional feeling over time requires extensive training data, with annotations typically performed by multiple annotators for reliability, which is both difficult and expensive to acquire.

Recently, *crowdsourcing* (CS) has become popular for performing tedious tasks through extensive human collaboration via the Internet. When it is difficult to employ experts for analyzing large-scale data, CS is an attractive alternative, as many individuals work on smaller data chunks to provide useful information in the form of annotations or tags. CS has been successfully employed to develop data-driven solutions for computationally difficult problems in multiple domains like natural language processing [30], and computer vision [33]. Two reasons mainly contribute to the success of CS– (1) crowd workers are paid a fraction of the wages that experts are entitled to, thereby achieving cost efficiency, and (2) the experimenter's task becomes scalable when the original task is split into smaller and manageable micro-tasks and distributed among crowdworkers. Nevertheless, cost-effectiveness in CS is achieved at the expense of expertise– crowdworkers may lack the technical and cognitive skills or the motivation to effectively perform a given task [19]. Therefore, efficient methodologies that are robust to noisy data are crucial to the success of CS approaches.

In this paper, we propose Multi-task learning (MTL) for time-continuous valence, arousal (VA) estimation from movie scenes for which dynamic emotion annotations are acquired from crowdworkers. Given a set of *related* tasks, MTL seeks to *simultaneously* learn all tasks by modeling the similarities as well as differences among them to build task-specific classification or regression models. This joint learning procedure accounting for task relationships leads to more efficient models as compared to learning each task independently. For the purpose of learning the relationship between low-level audio-visual features and corresponding crowdworker VA annotations over time, we ask the following questions: (1) Given that emotion perception is highly subjective, and biases relating to crowdworker demographics may additionally exist, can we discover any patterns relating to their dynamic emotional perception? The exercise of seeking to acquire a ***gold standard*** anno-

CrowdMM'14, November 7, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2660114.2660126.

tation for each movie scene (or clip) via crowdsourcing is meaningful only if such patterns can be discovered. (2) If the emotional ground truth corresponding to each movie clip can be represented by a single, gold standard emotional profile, can we discover corresponding audio-visual correlates for a movie clip collection (as against a single movie clip), which in turn, can be more effective for predicting the VA profile of a novel clip? Through extensive experiments, we demonstrate that MTL effectively answers the above questions, and is superior to single-task learning for VA prediction in novel scene segments. To summarize, this paper makes the following contributions:

1. This is the first work to employ MTL for time-continuous emotion prediction.
2. This is also one the first work to attempt dynamic affect prediction for movie clips.

The paper is organized as follows: Section 2 overviews the literature. Experimental protocol employed for recording crowdworkers' affective responses is described in Section 3 and a brief overview of MTL is provided in Section 4. Annotation data analysis and emotion prediction experiments are presented in Section 5, and conclusions are stated in Section 6.

2. RELATED WORK

We now examine related work on (1) Crowdsourcing, (2) Affective movie analysis, (3) CS for affective media tagging and (4) Multi-task learning.

2.1 Crowdsourcing

Steiner *et al.* [25] defined three types of video events and showed that these events can be detected from video sequences via crowdsourcing upon combining textual, visual and behavioral cues. Vondrick *et al.* [28] argued that frame-by-frame video annotation is essential for a variety of tasks, as in the case of time-continuous emotion measurement, even if it is difficult for human annotators. An online framework to collect valid facial responses to media content was proposed in the work of McDuff *et al.* [16], who found significant differences between subgroups who liked/disliked or were familiar/unfamilar with a particular commercial.

2.2 Affective movie analysis

A primary issue in affective multimedia analysis is the paucity of reliable annotators to generate sufficient training data and in most studies, only few annotators are used [21, 29]. Also, emotion perception varies with individual traits such as personality [10], and significant differences may be observed in affective ratings compiled from different persons over a small population. To address this problem, a number of studies have turned to crowdsourcing. In a seminal study affective movie study, Gross *et al.* [6] compiled a collection of movie clips to evoke eight emotional states such as anger, disgust, fear and neutral based on emotion ratings compiled for 250 movie clips from 954 subjects. [2, 11, 24] are three recent works that have attempted affect recognition from physiological responses of a large population of users to music and movie stimuli.

2.3 Crowdsourcing for affective media tagging

Soleymani *et al.* [22] performed crowdsourcing on a limited scale to collect 1300 affective annotations from 40 volunteers for 155 Hollywood movie clips. In another CS-based affective video annotation study, Soleymani *et al.* [23] compiled annotations for the MediaEval 2010 Affect Task Corpus on AMT, and asked workers to self-report their boredom levels. In a recent CS-based media tagging work, Soleymani *et al.* [20] presented a dataset of 1000 songs for music emotion analysis, each annotated continuously over time by at least 10 users. Nevertheless, movies denote multimedia stimuli that best approximate the real world and movie clips have been found to be more effective for eliciting emotions in viewers as compared to music video clips in [1], and that is why we believe continuous emotion prediction with movie stimuli is important in the context of affective media representation and modeling.

Table 1: Video clip details. HALV, LALV, HAHV and LALV respectively correspond to high-arousal low-valence, low-arousal low-valence, high-arousal high-valence and low arousal-low valence labels.

	HALV	LALV	HAHV	LAHV
No. of video clips	3	3	3	3
Min. length (sec)	79	80	86	59
Max. length (sec)	91	121	109	92
Avg. length (sec)	86.66	97.33	101	76.33

2.4 Multi-task learning

Recently, multi-task learning (MTL) has been employed in several computer vision applications such as image classification [32], head pose estimation [31] and visual tracking [34]. Given a set of related tasks, MTL [4] seeks to simultaneously learn a set of task-specific classification or regression models. The intuition behind MTL is simple: a joint learning procedure which accounts for task relationships is expected to lead to more accurate models as compared to learning each task separately. While MTL has been used previously for learning from noisy crowd annotations [9], we present the first work that employs MTL for time-continuous emotion prediction from movie clips.

3. EXPERIMENTAL PROTOCOL

In this study, we asked crowd workers to continuously annotate 12 emotional movie scenes adopted from [1] via a web-based user interface– they were not allowed to access the scene content prior to the rating task. Our objective was to understand and model their emotional state over time, as they viewed the movie clips.

3.1 Dataset

We selected 12 video clips from [1] equally distributed among the four quadrants in the valence-arousal space. Table 1 presents characteristics of video clips from the different quadrants. All video clips were hosted on YouTube for access during the CS task.

3.2 Experimental Protocol

We posted the annotation task on Amazon Mechanical Turk (AMT) and other CS channels via the CrowdFlower (CF) platform. CF is an intermediate platform for posting the AMT task on our behalf. Moreover, CF provides a simple gold standard qualification mechanism to discard outliers. If workers passed the qualification test, they were considered qualified to perform a given task. However, pre-designed tests are very generic and limited to simple tasks, which do not allow for trivially discarding low quality annotations. So, we performed PHP server-side scripting and redirection, collection and evaluation of all annotations real-time on our server via HTTP requests, before letting workers submit the task. The architecture of the designed CS platform is shown in Fig. 1(a).

To ensure annotation quality, each crowd worker could only annotate 5 video clips, and at least 15 judgments were requested and collected for each video clip. We also recorded facial expressions

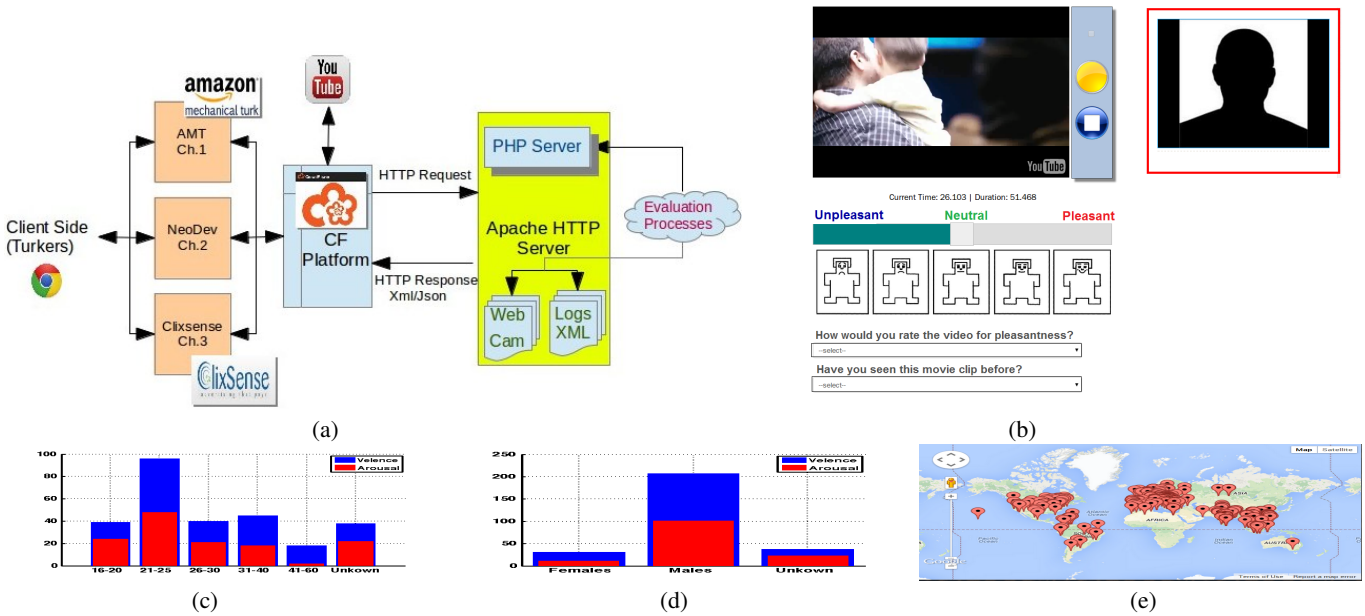

Figure 1: (a) Architecture of the designed Crowdflower platform (Turkers are the AMT crowdworkers). (b) User-interface for recording workers' emotional ratings and facial expressions. (c) Age, (d) gender and (e) locality distributions of crowdworkers.

of workers (not used in this work) as they performed the annotation. Informed consent was obtained from workers and, prior to the task, workers had to provide their demographics (age, gender and location). Time-continuous valence/arousal annotations from workers were compiled over separate sessions (a worker need not annotate for both valence and arousal for the same clip under this setting), and workers were also required to rate each clip for overall emotional valence or arousal. Each worker was paid 10 cents per video as remuneration upon successful task completion.

Workers did not get paid if their annotations and webcam facial videos were not recorded on our server. To evaluate the annotation quality, each video annotation was logged in XML format and analyzed. A continuous slider was used to record emotional rating, and if the slider had not moved for more than 80% of the clip duration, or if more than 20% of the data was lost, the annotation was automatically discarded. Also, files smaller than a pre-defined threshold were discarded. If the annotation task was left incomplete, a warning message notified the worker about the missing annotations. Workers could then re-annotate the missing videos and get paid. For motivating workers to provide good quality annotations, we rewarded them with online gift vouchers if they provided high-quality annotations. Furthermore, we introduced some constraints such as: (1) Workers could not play (or rate) multiple video clips simultaneously. (2) Workers could annotate a video as many times as they wanted to. (3) Workers were allowed to use only the Chrome browser for annotation due to unavailability of HTML5 technology support in other browsers. (4) Media player controllers were removed from the interface so that workers could not fast forward/rewind the movie clips, and finally, (5) If the annotation stopped midway, it had to be redone from scratch.

3.3 Annotation Mechanism

A screen shot of the user interface for recording annotations is presented in Fig. 1(b). The following components were part of the continuous annotation and facial expression recording process.

Video Player: To provide an uninterrupted video stream for workers with low bandwidth, we uploaded all movie clips onto YouTube. On the client-side, YouTube JavaScript player API was integrated and used in our web-based user interface.

Slider: The slider was used to collect the time-continuous VA ratings of workers while watching the video clips. The slider values ranged from -2 to 2 (*very unpleasant* to *very pleasant* for valence, and *calm* to *highly excited* for arousal) for both factors. In order to facilitate workers' decision making, a standard visual scale Self Assessment Manikin (SAM) image was displayed to the workers.

Webcam Panel: To upload facial expression of workers in real-time, we used HTML5 technology to buffer the worker's webcam recording on the client-side when the play button was pressed. The buffered video was automatically uploaded in compressed, VP8 open codec format on our server when the video clip finished playing. Videos were recorded at 320x240 resolution, 30 fps.

Questionnaires: Workers needed to report (1) their overall emotional (valence or arousal) rating for the movie clip on a scale of -2 to 2, and (2) their familiarity with the clip to avoid the effect of such bias on their ratings.

3.4 Annotation statistics and pre-processing

Overall, 1012 and 527 workers provided continuous valence and arousal ratings respectively. Their age, gender and locality distributions are as shown in Fig. 1(c),(d) and (e). As a preliminary step towards ensuring good quality VA labels, we discarded those time-continuous annotations with (1) missing values more than threshold, (2) standard deviation less than threshold, and (3) missing overall or general VA ratings.

4. MULTI-TASK LEARNING

As mentioned previously, Multi-task learning (MTL) models both similarities as well as differences among a set of related tasks, which is more beneficial as compared to learning task-specific models. Given a set of tasks $t = 1..T$, with $X(t)$ denoting training data for the task t and $Y(t)$ their corresponding labels (ratings), MTL seeks to jointly learn a set of weights $W = [W_1..W_T]$, where W_t models task t. For the problem of time-continuous VA prediction, the 12 movie clips used for crowdsourcing denote the related tasks.

In this work, we used the publicly available MALSAR library [35], which contains a host of MTL algorithms for analysis. We were particularly interested in the following MTL variants:

Multi-task Lasso: which extends the Lasso algorithm [26] to MTL, and assumes that sparsity is shared among all tasks.

ℓ_{21} **norm-regularized MTL [3]:** which attempts to minimize the objective function $\sum_{t=1}^{T} \|W_t^T X_t - Y_t\|_F^2 + \alpha\|W\|_{2,1} + \beta\|W\|_F^2$, where $\|.\|_F$ and $\|.\|_{2,1}$ denote matrix Frobenious norm and ℓ_{21} norm respectively. The basic assumption in this model is that *all* tasks are related, which is not always true, and α, β denote the regularization parameters controlling group sparsity and norm sparsity respectively.

Dirty MTL [8]: where the weight matrix $W = P + Q$, where P and Q denote the group and task-wise sparse components.

Sparse graph Regularization (SR-MTL): where *a priori* knowledge concerning task-relatedness is modeled in terms of a graph R in the objective function. This way, similarity is only enforced between W_t's corresponding to related tasks. The minimized objective function in this case is $\sum_{t=1}^{T} \|W_t^T X_t - Y_t\|_F^2 + \alpha\|WR\|_F^2 + \beta\|W\|_1 + \gamma\|W\|_F^2$, where R is the graph encoding task relationships, and α, β, γ denote regularization parameters as above.

5. DATA ANALYSIS AND EXPERIMENTS

Upon compiling VA ratings from crowdworkers, we firstly examined if any patterns existed in the dynamic annotations. This examination was important for two reasons– (1) predicting dynamic VA levels for a stimulus instead of the overall rating is useful as it allows for determining the 'emotional highlight' in the scene, and comparing dynamic vs static ratings could help us understand how dynamic emotion perception influenced crowdworkers' overall impression of a scene, and (2) given the subjectivity associated with emotions and the uncontrolled worker population, patterns in dynamic VA annotations would indicate that a reliable, *gold standard* annotation for a clip is achievable in spite of these biases.

Given the 12 clips (tasks) related in terms of valence and arousal (from now on, clips/tasks 1-3, 4-6, 7-9 and 10-12 respectively correspond to HAHV, LAHV, LALV and HALV labels), we employed MTL to determine *if some time-points influenced the overall emotional perception for a movie clip more than others*? To this end, we used the time-continuous VA ratings to *predict* the overall VA rating for each clip. For dimensional consistency, we only used the VA ratings for the final 50 sec of each clip for this experiment, and so, the x-axis in Fig. 2 denotes time to clip completion (between 50-1 sec). Weights learnt using the different MTL variants consistently suggest that the continuous VA ratings provided in the *latter* half for ***all*** of the movie scenes predict the overall rating better. The third and fourth columns respectively depict learnt weights for the six HV, LV/HA, LA clips, and represents the situation where *a-priori* knowledge regarding task-relatedness is fed to SR-MTL. Examining SR-MTL valence weights (cols 3,4 in row 1), one can infer that general affective impressions are created earlier in time for high-valence stimuli as compared to low-valence stimuli. Examination of SR-MTL arousal weights suggests that the most influential impressions regarding high-arousal stimuli are also created a few seconds before clip completion.

Therefore, MTL enables effective characterization of patterns concerning dynamic VA levels of crowdworkers, and this in turn implies that deriving a representative, gold standard annotation from worker annotations for each movie clip is meaningful. While MTL has been used to learn from noisy crowd data [9], we simply used the median value of the annotations at each time-point to derive the ground truth emotional profile for each movie clip. Next, we will briefly describe the audio-visual features extracted from each clip, and show how the joint learning of the relationship between audio-visual features and VA ratings allows for more effective dynamic emotion prediction.

5.1 Multimedia Feature Extraction

Inspired by previous affective studies [18,29], we extracted low-level audio-visual features that have been found to correlate well with the VA dimensions. In particular, we extracted the features used in [7,12] on a per-second basis for our regression experiments.

5.1.1 Video Features

Lighting key and color variance [29] are well-known video featuresknown to evoke emotions. Therefore, we extracted lighting key from each frame in the HSV space by multiplying the mean by the standard deviation of V values. Color variance [12] is defined as the determinant of the covariance matrix of L, U, and V in the CIE LUV color space. Also, the amount of motion in a movie scene is indicative of its excitement level [12]. Therefore, we computed the optical flow [15] in consecutive frames of a video segment to motion magnitude for each frame. The proportions of colors are important elements for evoking emotions [27]. A 20-bin color histogram of hue and lightness values in the HSV space was computed for each frame of a segment and averaged over all frames. The mean of the bins reflect the variation in the video content. For each frame in a segment, the median of the L and S values in HSL space were computed; their average for all the frames of a segment is an indication of the segment lightness and saturation [12]. We also used the definitions in [29] to calculate shadow proportion, visual excitement, grayness and visual detail. Extracted video features are listed in Table 2.

5.1.2 Audio Features

Sound information in the form of loudness of speech (energy of sound) is related to arousal, while rhythm and average pitch in speech relates to valence [18], while Mel-frequency cepstrum components (MFCCs) [14] are representative of the short-term sound power spectrum. Commonly used features in audio and speech processing [14] were extracted from the audio channels. To extract MFCCs, we divided the audio segment into 20 divisions and then extracted the first 13 MFCC components from each division. Using the sequence of MFCC components over a segment, we computed 13 derivatives of MFCC, DMFCC, and mean auto correlation, AMFCC proposed in [14]. Upon calculating MFCC, DMFCC and AFCC (13 values each), we used their means as features. The implementation in [17] was used to extract formants up to 4400Hz over the audio segment, and formant means were used as features. Moreover, we used the ACA toolbox [13] to calculate mean and standard deviation(std) of (i) spectral flux, (ii) spectral centroid and (iii) time-domain zero crossing rate [14] over 20 audio segment divisions. We also calculated the power spectral density and the bandwidth, band energy ratio (BER) , and density spectrum magnitude (DSM) according to [14]. Finally, we also computed the mean proportion of silence as defined in [5]. All in all, 56 audio features listed in Table 2 were extracted.

5.2 Experiments and Results

In this section, we attempt to *predict* the gold standard (or ground-truth) dynamic V/A ratings for each clip from audio-visual features using MTL, and show why learning the audio visual feature-

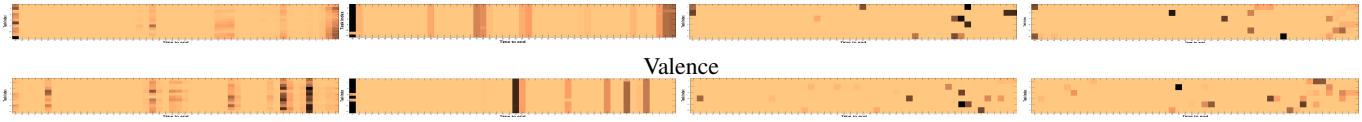

Figure 2: Predicting overall clip emotion from dynamic annotations: (Top) W matrix learnt for valence using (from left to right) $\ell_{2,1}$, dirty and SR MTL (HV, LV). (Bottom) W learnt for arousal using (from left to right) $\ell_{2,1}$, dirty and SR MTL (HA, LA). Larger weights are denoted using darker shades.

Table 2: Extracted audio-visual features from each movie clip (feature dimension listed in parenthesis).

Audio features	Description
MFCC features (39)	MFCC coefficients [14], Derivative of MFCC, MFCC Autocorrelation (AMFCC)
Energy (1) and Pitch (1)	Average energy of audio signal [14] and first pitch frequency
Formants (4)	Formants up to 4400Hz
Time frequency (8)	mean and std of: MSpectrum flux, Spectral centroid, Delta spectrum magnitude, Band energy ratio [14]
Zero crossing rate (1)	Average zero crossing rate of audio signal [14]
Silence ratio (2)	Mean and std of proportion of silence in a time window [5, 14]
Video features	**Description**
Brightness (6)	Mean of: Lighting key, shadow proportion, visual details, grayness, median of Lightness for frames, mean of median saturation for frames
Color Features (41)	Color variance, 20-bin histograms for hue and lightness in HSV space
VisualExcitement (1)	Features as defined in [29]
Motion (1)	Mean inter-frame motion [15]

emotion relationship simultaneously for the 12 movie scenes is more effective than learning scene-specific models. Fig. 3 shows the model weights learnt by the various MTL approaches when they are trained with features and VA ratings over the entire clip duration for all clips. Here again, some interesting correlates between audio-visual features and VA ratings are observed over all scenes. Considering video features, color descriptors are found to be salient for valence, while motion and visual excitement correlate with arousal better, especially for HA stimuli, as noted from SR MTL (HA) weights (column 7). Among audio features, the first few MFCC components correlate well with both V,A.

Then, we examined if learning prediction models for all movie clips was more beneficial than training a Lasso regressor per movie clip. To this end, we held out time-contiguous data of length 5, 10 or 15 seconds from the first half (front) or second half (back) of each of the clips for testing, while the remainder of the clips were used for training. Optimal group sparsity regularization parameter for the different MTL methods, as well as optimal Lasso parameter were chosen from [0.01 0.1 1 5] employing 5-fold cross validation, and all other parameters (where necessary) were set to 1. The root mean square error (RMSE) observed for V/A estimates over all clips (tasks) is shown in Table 3. MTL methods clearly outperform single-task Lasso, and consequent to our earlier finding that the latter half of all clips is emotionally salient, larger prediction errors are observed for the back portion. Also, prediction errors increase with the test clip size, and predictions are more accurate for arousal, and with audio features. Finally, sophisticated MTL methods such as dirty and SR-MTL outperform MT-Lasso and $\ell_{2,1}$ MTL. Overall, these results are demonstrative of efficient MTL-based learning utilizing relatively few training examples.

6. CONCLUSION AND FUTURE WORK

This paper explores Multi-task learning to estimate dynamic VA levels for movie scenes. Since time-continuous VA annotations are highly difficult to acquire, we employ crowdsourcing for the same. Though emotion is a subjective feeling and the crowdworkers arose from varied demographics, MTL could effectively capture patterns concerning their dynamic emotion perception .The latter half of all clips was found to be more emotionally salient, and influenced the affective impression of the clip. We again utilized MTL to model the relationship between the representative dynamic VA profile for each clip and underlying audio-visual effects, and observed that MTL approaches considerably outperformed clip-specific Lasso models, implying that jointly learning characteristics of a collection of scenes is beneficial. Future work involves usage of (1) MTL for cleaning crowd annotations, and (2) face videos compiled in this work as an additional affective cue.

7. ACKNOWLEDGEMENT

The authors acknowledge that this research was funded by the research grant for ADSC's Human Sixth Sense Programme from Singapore's Agency for Science, Technology and Research (A*STAR), the FIRB 2008 project S-PATTERNS, and cluster project "Ageing at Home".

8. REFERENCES

[1] M. Abadi, S. Kia, R. Subramanian, P. Avesani, and N. Sebe. User-centric affective video tagging from MEG and peripheral physiological responses. In *Affective Computing and Intelligent Interaction*, pages 582–587, 2013.

[2] M. K. Abadi, M. Kia, S. Ramanathan, P. Avesani, and N. Sebe. Decoding affect in videos employing the MEG brain signal. pages 1–6, 2013.

[3] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Neural Information Processing Systems*, 2007.

[4] R. Caruana. *Multitask learning*. Springer, 1998.

[5] L. Chen, S. Gunduz, and M. T. Ozsu. Mixed type audio classification with support vector machine. In *IEEE Int'l Conference on Multimedia and Expo*, pages 781–784, 2006.

[6] J. J. Gross and R. W. Levenson. Emotion elicitation using films. *Cognition & Emotion*, 9(1):87–108, 1995.

[7] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[8] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *Neural Information Processing Systems*, 2010.

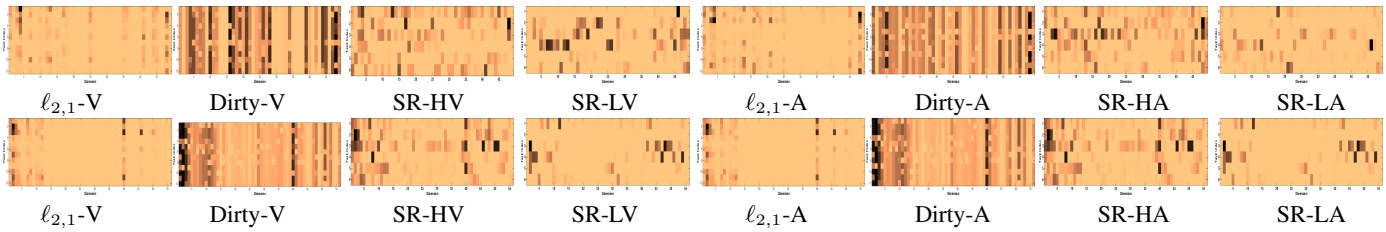

Figure 3: Predicting dynamic V/A ratings using (top) video and (bottom) audio features. Larger weights shown using darker shades (best viewed under zoom).

Table 3: RMSE-based V/A prediction performance of task-specific vs multi-task methods. RMSE mean, standard deviation over five runs are reported. Best model RMSE is shown in bold.

			Front			Back		
			5 s	10 s	15 s	5 s	10 s	15 s
Valence	**Video**	Lasso	0.429±0.041	0.816±0.583	1.189±0.625	0.584±0.024	0.881±0.057	1.125±0.064
		MT-Lasso	**0.191±0.028**	**0.319±0.064**	**0.549±0.042**	**0.206±0.014**	**0.443±0.067**	**0.593±0.108**
		ℓ_{21} MTL	0.193±0.030	0.326±0.063	0.565±0.047	0.207±0.015	0.450±0.066	0.606±0.113
		Dirty MTL	0.452±0.141	0.840±0.293	1.179±0.400	0.308±0.105	0.607±0.140	0.801±0.129
		SR MTL	0.193±0.030	0.325±0.064	0.563±0.046	0.207±0.015	0.450±0.066	0.607±0.113
	Audio	Lasso	0.475±0.030	0.712±0.069	0.851±0.081	0.634±0.016	0.860±0.027	1.174±0.034
		MT-Lasso	**0.241±0.023**	**0.348±0.014**	**0.487±0.038**	**0.237±0.024**	**0.400±0.039**	**0.527±0.033**
		ℓ_{21} MTL	0.243±0.020	0.359±0.012	0.520±0.029	0.247±0.026	0.392±0.029	0.553±0.028
		Dirty MTL	0.299±0.023	0.473±0.019	0.751±0.060	0.312±0.043	0.524±0.042	0.692±0.072
		SR MTL	0.248±0.017	0.365±0.015	0.526±0.027	0.252±0.027	0.404±0.033	0.567±0.026
Arousal	**Video**	Lasso	0.429±0.041	0.816±0.583	1.189±0.625	0.584±0.024	0.881±0.057	1.125±0.064
		MT-Lasso	**0.191±0.028**	**0.319±0.064**	**0.549±0.042**	**0.206±0.014**	**0.443±0.067**	**0.593±0.108**
		ℓ_{21} MTL	0.193±0.030	0.326±0.063	0.565±0.047	0.207±0.015	0.450±0.066	0.606±0.113
		Dirty MTL	0.452±0.141	0.840±0.293	1.179±0.400	0.308±0.105	0.607±0.140	0.801±0.129
		SR MTL	0.193±0.030	0.325±0.064	0.563±0.046	0.207±0.015	0.450±0.066	0.607±0.113
	Audio	Lasso	0.435±0.038	0.599±0.050	0.727±0.058	0.556±0.033	0.807±0.037	1.004±0.033
		MT-Lasso	**0.212±0.026**	**0.339±0.020**	**0.406±0.019**	0.243±0.039	**0.353±0.028**	**0.464±0.029**
		ℓ_{21} MTL	0.212±0.028	0.345±0.022	0.437±0.027	**0.238±0.032**	0.358±0.022	0.475±0.027
		Dirty MTL	0.249±0.015	0.420±0.036	0.618±0.051	0.304±0.043	0.430±0.029	0.568±0.019
		SR MTL	0.214±0.027	0.350±0.031	0.431±0.026	0.242±0.031	0.363±0.026	0.487±0.030

[9] H. Kajino, Y. Tsuboi, and H. Kashima. A convex formulation for learning from crowds. In *AAAI Conference on Artificial Intelligence*, 2012.

[10] E. G. Kehoe, J. M. Toomey, J. H. Balsters, and A. L. W. Bokde. Personality modulates the effects of emotional arousal and valence on brain activation. *Social Cognitive & Affective Neuroscience*, 7:858–70, 2012.

[11] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[12] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[13] A. Lerch. *An introduction to audio content analysis: Applications in signal processing and music informatics*. John Wiley & Sons, 2012.

[14] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22(5):533–544, 2001.

[15] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *Int'l Joint Conference on Artificial Intelligence*, volume 81, pages 674–679, 1981.

[16] D. McDuff, R. Kaliouby, and R. W. Picard. Crowdsourcing facial responses to online videos. *IEEE Transactions on Affective Computing*, 3(4):456–468, 2012.

[17] K. Mustafa and I. C. Bruce. Robust formant tracking for continuous speech with speaker variability. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(2):435–444, 2006.

[18] R. W. Picard. *Affective computing*. MIT press, 2000.

[19] J. Ross, L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Human Factors in Computing Systems*, pages 2863–2872, 2010.

[20] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang. 1000 songs for emotional analysis of music. In *ACM international workshop on Crowdsourcing for multimedia*, pages 1–6, 2013.

[21] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun. Affective characterization of movie scenes based on multimedia content analysis and user's physiological

emotional responses. In *IEEE Int'l Symposium on Multimedia*, pages 228–235, 2008.

[22] M. Soleymani, J. Davis, and T. Pun. A collaborative personalized affective video retrieval system. In *Affective Computing and Intelligent Interaction*, pages 1–2, 2009.

[23] M. Soleymani and M. Larson. Crowdsourcing for affective annotation of video: development of a viewer-reported boredom corpus. In *Workshop on Crowdsourcing for Search Evaluation*, 2010.

[24] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *T. Affective Computing*, 3(1):42–55, 2012.

[25] T. Steiner, R. Verborgh, R. Van de Walle, M. Hausenblas, and J. G. Vallés. Crowdsourcing event detection in youtube video. In M. Van Erp, W. R. Van Hage, L. Hollink, A. Jameson, and R. Troncy, editors, *Workshop on detection, representation, and exploitation of events in the semantic web*, pages 58–67, 2011.

[26] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[27] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394, 1994.

[28] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *Int'l Journal of Computer Vision*, 101(1):184–204, 2013.

[29] H. L. Wang and L.-F. Cheong. Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(6):689–704, 2006.

[30] J. D. Williams, I. D. Melamed, T. Alonso, B. Hollister, and J. Wilpon. Crowd-sourcing for difficult transcription of speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 535–540, 2011.

[31] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, and N. Sebe. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *IEEE Int. Conf. on Computer Vision*, pages 1177–1184, 2013.

[32] X. Yuan and S. Yan. Visual classification with multi-task joint sparse representation. In *Computer Vision and Pattern Recognition*, 2010.

[33] J. Yuen, B. C. Russell, C. Liu, and A. Torralba. Labelme video: Building a video database with human annotations. In *Int'l Conference on Computer Vision*, pages 1451–1458, 2009.

[34] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via structured multi-task sparse learning. *Int'l Journal of Computer Vision*, 101(2):367–383, 2013.

[35] J. Zhou, J. Chen, and J. Ye. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University, 2011.