Multimodal Detection of Salient Behaviors of Approach-Avoidance in Dyadic Interactions

Bo Xiao[†], Panayiotis G. Georgiou[†], Brian Baucom[‡], Shrikanth S. Narayanan[†,‡]
† Department of Electrical Engineering, ‡ Department of Psychology
University of Southern California, Los Angeles, CA 90089
boxiao@usc.edu, georgiou@sipi.usc.edu, baucom@usc.edu, shri@sipi.usc.edu

ABSTRACT

Approach-Avoidance (AA) coding is a measure of involvement and immediacy in human dyadic interactions. We focus on analyzing the salient events in interactions that trigger change points in AA code in time, as perceived by domain experts. We employ coarse level visual cues associated with body parts, as well as vocal energy features. Motion vector extraction and body pose estimation techniques are used for extracting visual cues. Functionals of these cues are used as features for SVM based machine learning experiments. We found that the coder's judgments on salient events are related to the short time interval preceding the labeling. We also show that visual cues are the main information source for decision making on salient AA events, and that considering the information from a subset of body parts provides the same information as considering the full set. The mean of absolute value and standard deviation of motion streams are the most effective functionals as feature. We achieve an F-score of 0.55 in detecting salient events using cross-validation with a one-subject-out approach.

Categories and Subject Descriptors

J.4 [**Social And Behavioral Sciences**]: Psychology

General Terms

Algorithms

Keywords

Approach-Avoidance, salience, pose estimation, motion vector, multimodal feature

1. INTRODUCTION

The computational analysis of human interactions opens up new challenges and opportunities for signal processing [1]. As one of the common interaction forms, dyadic interactions take place in everyday life as well as in particular application scenarios such as psychotherapy and diagnosis, doctor-patient interviews, and customer care. Psychologists have been studying human behavior including notably in research and clinical settings and developed various coding methods, e.g. [2], as instruments to quantitatively describe behavior in domain-specific dimensions. The ability to model the relation between multimodal observations and expert-derived codes is valuable for both the fields of engineering and psychology, in terms of offering processing efficiency and scalability as well as enabling new tools for seeking insights for the domain experts.

There are mainly two levels of behavior coding. High level codes are often given to entire interaction sessions which are more abstract and summative of constituent behavior details. For example, Black *et al.* [3] and Georgiou *et al.* [4] demonstrated automatic classification of couples' behavior in therapy sessions. In a 96-hour corpus, coders assigned various high level codes to each spouse, such as "acceptance", "blame", *etc.* By using a combination of automatically-derived lexical information and speech features, they demonstrated prediction of the behavior codes with accuracy exceeding 80%.

On the other hand, intermediate level codes are often specified moment by moment, and capture dynamic aspects of the behavior. Oertel *et al.* studied involvement in a group conversation [5]. Two 30 min segments with 4 or 5 subjects were used in their analysis. A one dimensional code along the time axis was assigned to all the participants as an average degree of involvement. Using multimodal features including prosodic and visual cues such as gaze and blinking, they achieved an accuracy of 68% in classification of three involvement levels with a SVM.

Approach-Avoidance (AA) [6] is another intermediate level behavior related to involvement and immediacy that has been studied extensively. It is typically assigned on an individual basis continuously in time using a numerical rating, e.g., on a 9-point likert scale from -4 to 4. The AA code is associated with subject behaviors that can be directly observed such as head, hand and body pose, movement, gaze, facial expression, speech fluency and latency, *etc*, hence suggesting that the use of signal processing is an appropriate tool for its estimation. Rozgić *et al.* [7] studied the problem of estimating the AA code in an Ordinal Logistic Regression (OLR) framework. Acoustic features and motion features derived from a Vicon Motion Capture system were considered. The functionals of these features in time domain were used as input for OLR. By combining OLR with an HMM model to capture dynamics, the algorithm could estimate the AA code with accuracy of 75%. In addition,

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.

their results showed that the AA behavior was more robustly estimated using the visual cues.

While the past work has focused on estimating the code value, not much has been done in predicting the dynamics of AA code change. In fact, over the course of an interaction, the information encoded and conveyed typifying certain behavior patterns are not continuously and uniformly manifested over time. There are salient events which "highlight" specific behavior patterns. Psychologists also refer to the phenomenon of "thresholding" negative influence from the other interlocutor in conversations [8], where people would typically show behaviors whenever the negative influence exceeds a certain tolerance, rather than react to the influence immediately. The moment that the subject makes some substantial change in his/her behavior in response is considered salient. Therefore, computationally modeling these salient phenomena of human behavior is also important. In addition, human-agreement on absolute degree of a behavior code is often arguably lower than agreement on the relative salience.

In this paper, we investigate temporal changes in the AA code, that may indicate a salient event in the interaction. We focus on the coarse level body movements as they are the most prominent observable phenomena. Pixel-wise motion vectors are computed per frame and associated with estimated body parts, which is then summed over blocks and used as the visual cue. Voice energy is employed as an acoustic cue. For a given time point, we analyze a short time interval preceding that point. Functionals of the visual and acoustic cues are adopted as features for classification and detection of salient events. We introduce the data set in Sec. 2, explain the feature extraction procedure in Sec. 3, report experimental results in Sec. 4 and conclude the work in Sec. 5.

2. DATA SET

For the experiments in this paper, we used the multimodal dyadic interaction database presented in [9]. The sessions in this corpus were dyadic interactions based on unscripted role-playing on conflictual topics, such as cheating in relationship, arguing over a drinking problem, *etc.* In order to get interactions that are as realistic as possible, participants were provided with the topics a few days before the recording. At the time of collection, participants were allowed several minutes to prepare for a topic, where they could exchange ideas and make up a story. During the collection, the two subjects took seats on a couch side by side.

The corpus provides near-field frontal view video data for each subject in resolution of 1024×768 and 30 frames per second. Audio data were collected from close-talking lapel microphones. One psychologist annotated eight subjects involved in 32 sessions with respect to individual subject. The total length of the recording is about 135 minutes. The psychologist gave an Approach-Avoidance (AA) code that is discrete in value from -4 to 4, and continuous in time. During annotating the psychologist followed the "three seconds rule", which means a change of AA code is in general based on the observation of the last three seconds. The perceptually salient events are identified by a change of AA code whenever a new code value is assigned along time. When the code value remains stationary, we call that period as perceptually non-salient. In total we found 659 cases of salient events. The details of the data are shown in Table 1.

Table 1: Summary of data from each subject

Subject	1	2	3	4	5
No. session	4	2	6	6	4
Total length (min)	24	5	21	21	13
No. salient events	73	24	118	112	69

Subject	6	7	8	Total
No. session	4	3	3	32
Total length (min)	13	19	19	135
No. salient events	43	124	96	659

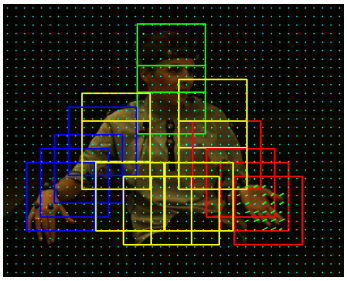


Figure 1: Body pose and motion vector estimation

3. MULTIMODAL FEATURE EXTRACTION

3.1 Motion vector and body pose

We extract the pixel-wise motion vector using Farnebäck's algorithm [10] that is implemented in OpenCV library on a frame basis. However, irrelevant motion may appear in the background. More importantly, the raw motion vectors lack the association with the parts of human body. To address this problem, we use the state of the art body pose estimation approach introduced in [11]. This approach captures orientation with a mixture of templates for each body part, and contextual co-occurrence relations as well as standard spring model encoded spatial relations between parts. When these relations are in a tree structure, the model can be efficiently optimized with dynamic programming.

In our work, we apply an extensively trained upper body model developed by the authors of the tool on each video frame, which returns 18 square blocks over the corresponding position, including 2 blocks for head, 4 blocks each for left and right arms, and 8 blocks for torso. As the two subjects were seated side by side, they talked to counter directions, which may lead to a swapped pattern of left and right arm behaviors. So for the interlocutor seated on the right, we swap the order of the left arm and torso blocks with the right ones to get features aligned for all subjects.

The original estimation procedure adopted an iterative optimization on the block size, which may vary among frames and deviate from the reasonable size. Since in our scenario, the subjects remain seated during the interaction, the block size did not vary significantly. Thus we decided to fix the block size by empirically evaluating the pose estimation using different sizes, and choosing the size that best fits the data. In Figure 1 we show an example of body pose and motion vector estimation.

3.2 Feature design

Let the time index of a video frame in a session be t. Let the row and column indexes of a pixel be i and j. The motion vector of a pixel (i, j) at time t is represented as

$(dx, dy) = f(i, j, t)$. Let $B(k, t)$ be the k-th block of body part at time t, and $M(k, t) = \sum_{(i,j)\in B(k,t)} f(i, j, t)$ be the corresponding raw motion stream. As a result, we obtain a 2-dimension motion stream for each body part, considering horizontal and vertical axes separately. In order to align acoustic and visual cues, we sample the energy of speech signal using the same frequency as video, i.e. 30 frames per second with no overlap. We obtain the vocal energy stream $V(t)$ as the last dimension of raw features.

In light of the "three seconds rule", we define a look-back window (LBW) of length W ending at a given time point t. We extract functionals of the raw feature streams as the final features. Functionals computed include mean of absolute value, standard deviation, absolute value of skewness, and kurtosis of the raw features. We also compute the second moment of the feature in the first and last third of the window, and take the larger of either their ratio or its reciprocal, in order to capture any temporal change of the feature stream. In addition, we compute the entropy of each feature stream. First, the raw feature stream is Z-normalized along the entire session, and then quantized to 8 bins with fixed boundaries. For each LBW the histogram of the quantized feature is used to compute the entropy. The functionals for all raw features are aggregated to get the final feature vector.

4. EXPERIMENTS

4.1 Experiment design

We sample one LBW per second in each session, with length W being 1, 2, 3 or 4 seconds. Normalization of feature value is applied to each subject to account for interpersonal behavior variability. The samples nearest to the salient events in time are labeled as in the salient class. The two closest neighboring samples of the salient sample are also considered as salient, in order to tolerate inaccuracy in coding as well as to cope with the time span of salient event. All other samples are labeled as in the non-salient class.

We found that direct modeling of salient and non-salient classes using all the samples did not yield good results. One reason might be that the salience property is not binary in nature, but thresholded by the coder at the time of perceiving the interaction. In addition the salient event might not be exact in time, but span a short period. Therefore the negative samples are contaminated by some partial salience. To overcome the problem, we use a balanced training design. Assume there are N coder-specified salient samples, then we select N non-salient samples that are at least 4 seconds away from each other as well as those coder-specified salient samples. This set is named as the balanced set. The linear SVM model is cross-validated by leaving one subject for testing at a time, and taking all other subjects for training.

We then apply the models derived with the balanced set to the entire sessions of the corresponding left-out subjects. In practice the size of non-salient class highly exceeds that of the salient class. To address this mismatch, we optimize the F-score of detecting salient events by tuning the threshold on the raw SVM output in making classification decision. Therefore we need a second layer of cross-validation, i.e. we leave one session of the subject for testing at a time, and take all other sessions as development set to carry out the optimization.

For evaluation, we report the accuracy on the balanced set, the precision, recall and F-score of detection on the entire dataset, as well as the Kappa value of agreement between the coder and our result, averaged with weighting on the sample size of each subject.

Table 2: Results of different LBW length

W (sec)	Acc.	Pre.	Rec.	F-score	Kappa
1	0.62	0.34	0.70	0.45	0.15
2	0.68	0.38	0.71	0.50	0.24
3	0.70	0.41	0.70	0.52	0.27
4	0.70	0.42	0.75	0.54	0.31

Table 3: Results by using different functionals

Functional	Acc.	Pre.	Rec.	F-score	Kappa
Mean abs.	0.68	0.42	0.76	0.54	0.31
Standard dev.	0.70	0.43	0.75	0.55	0.32
Skewness	0.60	0.31	0.82	0.45	0.10
Kurtosis	0.61	0.31	0.85	0.46	0.11
2nd moment ratio	0.57	0.32	0.76	0.44	0.12
Entropy	0.70	0.40	0.72	0.51	0.27

4.2 Experiment results

The results of different LBW length W are reported in Table 2. Note that the chance level for the "Acc" column is 0.5 as applied to the balanced set, while the chance level of F-score is 0.42, when claiming all samples as salient. Higher accuracy is obtained with longer LBW length. This suggests that the coder's perception of salience is dependent on a not too short period before the marked time point.

The performance of different functionals are reported in Table 3. W is set to 4 seconds in all of the following experiments. The most effective feature functionals are standard deviation and mean of absolute value. The entropy feature is close to the previous two. The higher order statistics and the second moment ratio do not result in good performance.

The results of using different multimodal cues are reported in Table 4, where all the functionals are incorporated. Here the row of a body part represents only using the motion cues of that part. The last row of "HSH" stands for a combination of upper head, two shoulders and two hands. The results show that using only the audio information does not effectively predict the salient events. The performance when using only some of the motion features does comparably with employing all the motion features. One reason might be that the movements of individual body parts are highly dependent due to the body's articulation. The other cause might be the body pose estimation. As shown in Figure 1, the blocks have a moderate degree of overlap, hence the motion of one body part is leaked to nearby parts. A finer and tighter body part estimation technique that can also resolve occlusion is therefore desired.

Based on the experiments above, we test the performance of selected functionals and cues, as shown in Table 5. MSE stands for mean of absolute value, standard deviation and entropy, while MS stands for the first two. V stands for selected visual features as the "HSH" row in Table 4, while A stands for acoustic features. We see that with only these features we achieve about the same performance as using all cues and functionals. Incorporating the entropy feature does not improve the result.

In Figure 2 we show one example of the precision-recall curve in the development phase, as well as the correspond-

Table 4: Results by using different multimodal cues

Cues	Acc.	Pre.	Rec.	F-score	Kappa
Audio	0.62	0.30	0.81	0.44	0.08
Motion	0.71	0.41	0.72	0.53	0.29
Head	0.69	0.40	0.78	0.53	0.27
Shoulder	0.67	0.39	0.78	0.52	0.26
Torso	0.68	0.40	0.70	0.51	0.26
Arm	0.69	0.43	0.69	0.53	0.31
Hand	0.69	0.42	0.68	0.52	0.29
HSH	0.72	0.42	0.77	0.54	0.31

Table 5: Results by using selective multimodal cues

Cues	Acc.	Pre.	Rec.	F-score	Kappa
MSE_AV	0.70	0.42	0.75	0.54	0.30
MS_AV	0.71	0.43	0.75	0.55	0.32

ing detection result on the left-out session. In general our approach suffers from low precision. This might be due to the limitation of coarse level motion and functionals. For example, not all motions are equally salient. Specific gestures and movements may bear more relevant behavioral meaning that is perceived as salient. Although the F-score is not very high in value, it is much higher than chance level. As a first attempt on the problem, the result shows the feasibility of using multimodal signal processing and machine learning approaches to infer salience in human dyadic interaction. We would also like to study human agreement as an upper bound when data with multiple annotation is available.

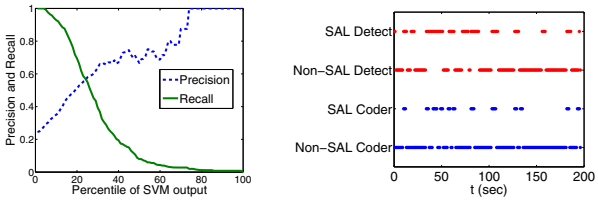
5. CONCLUSIONS

In this paper we focused on analyzing the salient events that trigger change points of the AA code, using coarse level visual cues and vocal energy features. State of the art motion vector extraction and body pose estimation were used for extracting visual cues. Functionals of visual and vocal cues were used as features in SVM based machine learning experiments. We found that the coder's perception of salient event is related to the short time interval preceding the change point of code. Visual cues are more informative in predicting salient events, while the subset of head, shoulder and hand yields similar results as the full feature set. The mean of absolute value and standard deviation of the motion streams are the most effective functionals. The best result of detecting salient events has an F-score of 0.55 using models learned on a balanced set and applied to entire sessions in a one-subject-out cross validation.

In the future we will pursue further work to improve the analysis and results with ongoing data collection. For example, better body part estimation techniques are desired. In terms of visual cues, one would like to incorporate finer details such as gaze and facial expressions. In order to separate meaning bearing gestures from casual movement, one needs to analyze the motion with a more complex time sequence model. Moreover, the salient events have their own properties such as the degree and direction of AA code change and we plan on modeling these properties.

6. ACKNOWLEDGMENTS

The research is supported in part by NSF, NIH, and DARPA.



(a) Precision and recall

(b) detection result

Figure 2: Example of detection result

7. REFERENCES

[1] A. Vinciarelli, H. Salamin, and M. Pantic. Social Signal Processing: Understanding social interactions through nonverbal behavior analysis. In *Proc. CVPR Workshops*, pages 42–49, 2009.

[2] C. Heavey, D. Gill, and A. Christensen. *Couples interaction rating system 2*. University of California, Los Angeles, 2002.

[3] M. Black, A. Katsamanis, C.C. Lee, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan. Automatic classification of married couples' behavior using audio features. In *Proc. InterSpeech*, 2010.

[4] P. Georgiou, M. Black, A. Lammert, B. Baucom, and S. Narayanan. "that's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features? In *Proc. ACII*, 2011.

[5] C. Oertel, S. Scherer, and N. Campbell. On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation. In *Proc. Interspeech*, 2011.

[6] L.K. Guerrero. Observer ratings of nonverbal involvement and immediacy. *The sourcebook of nonverbal measures: Going beyond words*, pages 221–235, 2004.

[7] V. Rozgic, B. Xiao, A. Katsamanis, B. Baucom, P. Georgiou, and S. Narayanan. Estimation of ordinal approach-avoidance labels in dyadic interactions: ordinal logistic regression approach. In *Proc. ICASSP*, 2011.

[8] J. Gottman, C. Swanson, and J. Murray. The mathematics of marital conflict: Dynamic mathematical nonlinear modeling of newlywed marital interaction. *Journal of Family Psychology*, 13(1):3–19, 1999.

[9] V. Rozgic, B. Xiao, A. Katsamanis, B. Baucom, P. Georgiou, and S. Narayanan. A new multichannel multimodal dyadic interaction database. In *Proc. InterSpeech*, 2010.

[10] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. *Image Analysis*, pages 363–370, 2003.

[11] Y. Yang and D. Ramanan. Articulated pose estimation using flexible mixtures of parts. In *Proc. CVPR*, 2011.