Multimodal Emotion Recognition in Response to Videos

통계학과 서석현

Introduction

- Affective self-reporting might be held in doubt, ~ … 자기 자신을 평가하는 것은 문제가 될 수 있다.
  -> 자기는 용감하다고 판단했지만 실제로는 겁에 질렸을 수 있다.

- multimedia 에 관련된 감정은 크게 3 가지로 연구되어 왔다.

1. Estimating emotions from multimedia content - Affective Video content representation and modeling (A. Hanjalic and L-Q. Xu, IEEE Trans Multimedia vol 7.)

2. Recognizing emotions induced by videos - Exploiting Facial Expressions for Affective Video Summerisation (H. Jobo, Proc ACM ~)

3. detect topical relevance or summarizing videos - Looking at the Viewer Analysing Facial Activity to detect personal highlights of multimedia Contents(H. Jobo, Multimedia Tools and applications)

- 논문에서는 EEG 신호와 Eye gaze 데이터를 사용해서 감정 인식을 했다.

- Irie et al은 토픽 모델링을 사용해서 영화의 음성 혹은 단어를 통해 감정을 인식하려고 했다.

감정 인식에 대한 대표적인 선행연구

- Using Noninvasive Wearable Computers to recognize Human Emotions from Physiological Signals (C.L. Lisette and F. Nasoz, Applied signal Processing vol 2004) -> 6개의 감정들을 동영상을 보고 Classification 진행함. 장점: 좋은 성능, 단점: 매우 감정적인 부분을 segment를 하여 보여줬다.

- Takahashi는 EEG를 통해 41.7의 정확도를 보이는 모델을 완성하였다. 하지만 feature level fusion 과 peripheral 한 신호는 성능 개선에 도움을 주지 못하였다.

- peripheral 시그널을 사용해 relevance vector machine을 사용한 Soleymani의 연구도 있었다.

- arousal, valence를 가지고 분류를 했더니 성능이 어느정도 좋아졌다.

- startle stimuli (random noise sounds)로 자극하고 분류를 하였더니 성능이 77.5로 좋아졌다.

- The Pupil as a Measure of Emotional Arousal and Autonomic Activation(M.M. Bradley, Psychophysiology vol 45) - 1000회 이상 인용

- Pupil size variation as an Indication of Affective Processing(T. partial, Human-computer studies, vol 59) - 600회 이상 인용

- Potential application은

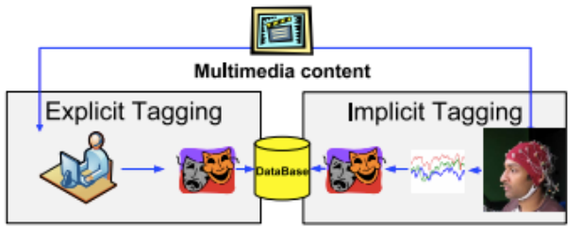


Fig. 1. Implicit affective tagging versus explicit tagging scenarios. The analysis of the bodily responses replaces the direct interaction between user and the computer. Therefore, users do not have to be distracted for tagging the content.

Material and Methods

- preliminary study에서는 2분 짜리 동영상을 155개 만들었다. 각 영상들에는 10개의 annotation들이 있었고 50명 이상의 피험자가 있었다.
- Emotion keyword(arousal, valence)에 대해서 9개 척도가 존재했다. 거기서 각각 많은 척도를 받은 14개 영상을 선택되었고 3개는 온라인에서(joy, happiness, disgust)와 3개는 기상일보 영상을 가져왔다.
- 영상 20개의 시간은 34.9 ~ 117초이며 M = 81.4, SD = 22.5초였다.
- facial video, audio, vocal expression, eye gaze, physicological signal이 녹화되었다.
- Tobii, Biosemi active 2 system
- ECG, EEG(32), galvanic skin response, respiration amplitude, skin temperature
- peripheral, vocal, facial X, EEG, pupillary response and gaze distance O
- 30명의 피험자 in Imperial College London, 17명의 여성 피험자, 13명의 남성 피험자, 19 ~40(m = 26.06, sd = 4.39), 다양한 배경
- 6명의 데이터는 품질이 좋지 못하여 24명으로 분석함. http://mahnob-db.eu

TABLE 1
The Video Clips and Their Sources

Code	Emotion Labels	Video clips sources
1	Act., Unp.	Hannibal
2	Act., Unp.	The Pianist
3	Med., Pls.	Mr. Bean's holiday
4	Act., Neu.	Ear worm (blip.tv)
5	Med., Neu.	Kill Bill VOL I
6	Med., Pls.	Love actually
7	Med., Pls.	Mr. Bean's holiday
8	Cal., Pls.	The thin red line
9	Med., Neu.	The shining
10	Med., Pls.	Love actually
11	Act., Unp.	The shining
12	Med., Unp.	Gangs of New York
13	Act., Unp.	Silent hill
14	Med., Unp.	The thin red line
15	Cal., Neu.	AccuWeather New York weather report (youtube.com)
16	Act., Unp.	American history X
17	Cal., Neu.	AccuWeather Detroit weather report (youtube.com)
18	Act., Pls.	Funny cats (youtube.com)
19	Cal., Neu.	AccuWeather Dallas weather report (youtube.com)
20	Act., Pls.	Funny (blip.tv)

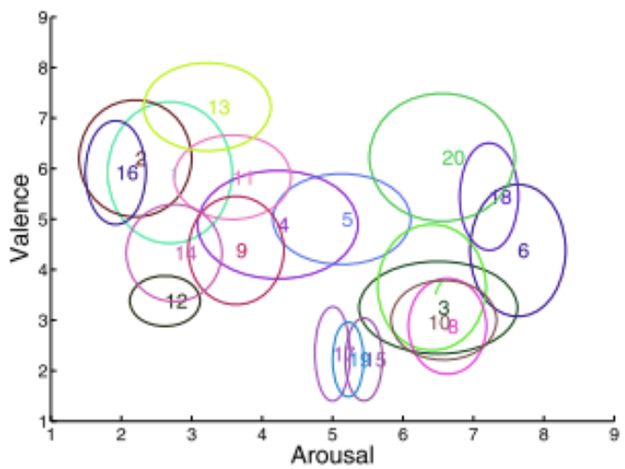


Fig. 3. Stimulus videos are shown in the valence-arousal plane. The center of the ellipses represents the mean arousal and valence and the horizontal and vertical radius represents the standard deviation of the online assessments. The clip codes are printed at the center of each ellipse.

Material and Methods

- 영상 상영 순서는 다음과 같고 특이한 점은 중간에 Neutral clip을 넣어서 감정을 환기하려고 노력함.
- Electroencephalogram signal, 1024Hz, 였지만 나중에 메모리 등 이유로 256Hz로 다운 샘플링함.
- preprocessing을 시켜서 noise를 제거함. 4-45 Hz band pass filter 사용
- 근육의 움직임을 최소한으로 하고자 함.
- full common mode rejection ratio(CMRR)를 50Hz로 얻기 위해 rereference to average reference to maximize signal to noise ratio.
- EEG in diffrent band가 감정과 상관관계가 있다.
- Power spectral density(PSD)를 다른 밴드에서 추출함? -FFT 알고리즘 등을 사용해서
- 겹치는 부분으로 신호를 분할, psd는 periodogram의 평균으로 예측. -> 좀 더 부드러운 power spectrum 15 s long window, 50 percent overlapping.
- theta (4, 8), slow alpha (8, 10), alpha, (8, 12), beta (12, 30), gamma (30,)
- the difference between left, right hemisphere 14개의 쌍 -> slow alpha를 제외한 나머지에서 asymmetry 한 모습을 보임
- 결국 변수 14 *4 + 32*5 = 216.

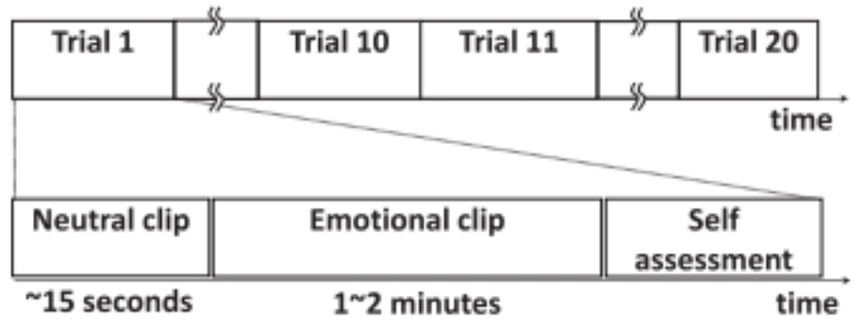


Fig. 5. Each trial started by a 15 s neutral clip and continued by playing one emotional clip. The self-assessment was done at the end of each trial. There were 20 trials in each session of the experiment.

TABLE 2
All the Features Extracted from Eye Gaze Data and EEG Signals

Eye gaze data	Extracted features
Pupil diameter	standard deviation, spectral power in the following bands:]0, 0.2]Hz,]0.2, 0.4]Hz,]0.4, 0,6]Hz and]0.6, 1]Hz
Gaze distance	approach time ratio, avoidance time ratio, approach rate
Eye blinking	blink depth, blinking rate, length of the longest blink, time spent with eyes closed
EEG	theta, slow alpha, alpha, beta, and gamma PSD for each electrode. The spectral power asymmetry between 14 pairs of electrodes in the four bands of alpha, beta, theta and gamma.

Material and Methods

- Eye gaze data from tobii 60Hz 사용
- a linear interpolation을 사용해서 눈이 깜박였을 때를 보정시킴
- 오른쪽, 왼쪽 눈의 diameter의 평균을 사용함.
- 빛에 대한 반응이 대부분이기 때문에 최대한 제거하려고 노력함.
- 빛에 대한 동공 반응에 대한 모형은 여러가지가 존재

$$Y = UDV^T, \quad (2)$$

$$A_p = UD, \quad (3)$$

$$S_p = V^T, \quad (4)$$

$$Y_1 = A_{p1}S_{p1}, \quad (5)$$

$$Y_R = Y - Y_1. \quad (6)$$

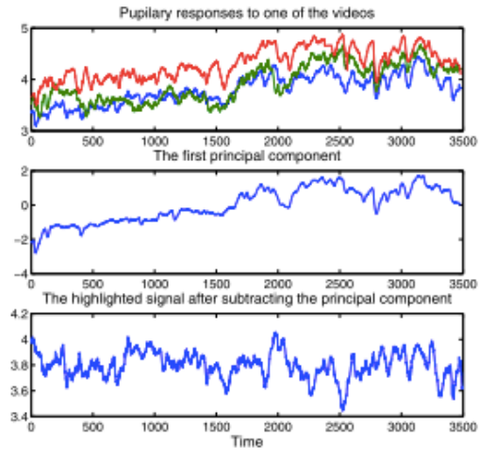


Fig. 7. From top to bottom: In the first plot, there is an example of pupil diameter measures from three different participants in response to one video. The second plot shows the first principal component extracted by PCA from the time series shown in the first plot (the lighting effect). The bottom plot shows the pupil diameter of the blue signal in the first plot after reducing the lighting effect.

1. photorealistic models for pupil light reflex and iridal pattern deformation(V.F pamplona, ACM trans Graphics vol28)
2. Modelling Autonomous Oscillations in the Human pupil light reflex using nonlinear delay-differential equations(A.longtin, Bull. math biology vol 51)

- 이러한 모형을 사용하는 것은 error 문제에 대해서 벗어날 수 없고 나이와 사람마다 편차에 영향을 너무 받는다.
- PCA를 통해서 lights reflex를 추정해보기로 함.
- Y = X + Z + E (Y = M * N_p, M은 centralized normalized pupillary response, N_p는 피시험자), X는 빛에 대한 반응, Z는 감정 혹은 집중, E는 에러.
- linear를 제거한 후 power spectrum을 적용함 - 동공동요(0.05~0.3Hz, 1mm 움직임) -> relaxed and passive, covering up to 0.4Hz.
- eye blink ~ anxiety, eye blinking rate, average, maximum of blink duration -> to detect unpleasant emotion.
- eye tracker의 distance를 통해서 실험자가 움직였나 안움직였나를 확인함. 실험이 계속될수록 사람이 자꾸 모니터 앞으로 가려고 함.
- These features were named approach and avoidance ratio to represent the amount of time each participant spent getting close to or going far from the screen. The frequency of the participants' movement toward the screen during each trial, approach rate, was also extracted

Emotion classification

- each video data set, using median arousal and valence score
- SVM classifier with RBF and loocv, feature selection은 one way ANOVA를 사용
- Modality fusion strategy (decision level, feature level)
- 본 논문에서는 EEG, eye gaze data를 통해서 decision level fusion을 진행함. 옆에 수식을 사용
-

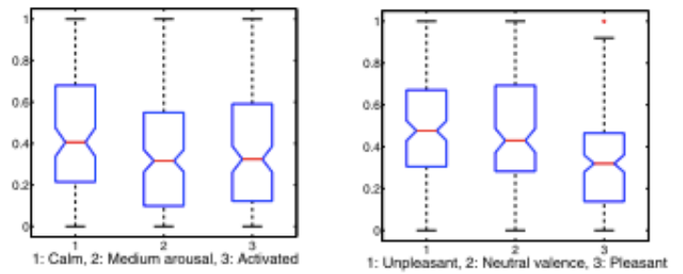
$$g_a = \frac{\sum_{q \in Q} P_q(\omega_a|x_i)}{\sum_{a=1}^{K} \sum_{q \in Q} P_q(\omega_a|x_i)} = \sum_{q \in Q} \frac{1}{|Q|} P_q(\omega_a|x_i). \quad (7)$$

In (7), g_a is the summed confidence interval for affect class ω_a. Q is the ensemble of the classifiers chosen for fusion, $|Q|$ the number of such classifiers, and $P_q(\omega_a|x_i)$ is the posterior probability of having class ω_a the sample is x_i according to classifier q. The final choice is done by selecting the class ω_a with the highest g_a. It can be observed that g_a can also be viewed as a confidence measure on the class, ω_a, given by the fusion of classifiers.

There are two problems in employing SVM classifiers in this fusion scheme. First, they are intrinsically only two-class classifiers and, second, their output is uncalibrated so that it is not directly usable as a confidence value in case one wants to combine outputs of different classifiers or modalities. To tackle the first problem, the one versus all approach is used where one classifier is trained for each class (N classifier to train) and the final choice is done by majority voting. For the second problem, Platt [49] proposes modeling the probability of being in one of the two classes knowing the output value of the SVM by using a sigmoid fit, while Wu et al. [50] propose a solution to extend this idea to multiple classes. In this study, we used the Matlab libSVM implementation [51] of the Platt and Wu algorithms to obtain the posterior probabilities, $P_q(\omega_a|x_i)$.

Result

- 24명의 피험자가 20개의 영상을 보았고 13개가 문제가 있었다.
- PCA를 적용했을 때 1st component가 50%의 var를 해석했다.
- Eye gaze는 ANOVA로 feature selection을 하였고
- EEG는 linear discriminantion criterion 사용
- cohen kappa는 0.32정도 나왔고 arousal 과 valence를 각 피험자 간 코릴레이션을 구해본 결과 arousal이 interrater간 유사하게 나옴.



(a) Eye blinking rate, arousal (b) Approach time ration, valence

(c) Blink depth, valence (d) SD of pupil diameter, valence

Fig. 8. Box plots of four different gaze data features in three emotional conditions. (a) Eye blinking rate for arousal classification. (b) Approach time ratio for valence classification. (c) Blink depth, average blink time, for valence classification. (d) STD of pupil diameter for valence classification. One way ANOVA results showed a significant difference between features mean of different classes ($p < 0.05$).

TABLE 3
Ten Best EEG Features for Arousal and Valence Classification Based on Linear Discrimination Criterion

Arousal classification			Valence classification		
Band	Electrode/s	$\sigma^2_{bw}/\sigma^2_{wn}$	Band	Electrode/s	$\sigma^2_{bw}/\sigma^2_{wn}$
Slow α	PO4	0.18	β	T8	0.08
α	PO4	0.17	γ	T8	0.08
θ	PO4	0.16	β	T7	0.07
Slow α	PO3	0.15	γ	T7	0.06
θ	Oz	0.14	γ	P8	0.05
Slow α	O2	0.14	γ	P7	0.05
Slow α	Oz	0.14	θ	Fp1	0.04
θ	O2	0.13	β	CP6	0.04
θ	FC6	0.13	β	P8	0.04
α	PO3	0.13	β	P7	0.04

The between class variance to within class variance ratios $\sigma^2_{bw}/\sigma^2_{wn}$ are also given.

Result

TABLE 4
The Classification Rate and F1 Scores of Emotion Recognition for Different Modalities

Modality	Classification rate		Average F1	
dimension	arousal	valence	arousal	valence
EEG	62.1%	50.5%	0.60	0.50
Eye gaze	71.1%	66.6%	0.71	0.66
Feature level fusion (FLF)	66.4%	58.4%	0.65	0.55
Decision level fusion (DLF)	76.4%	68.5%	0.76	0.68
Self-reports with SAM manikins	55.7%	69.4%	0.57	0.70
Random level	33.3%	33.3%	0.36	0.40

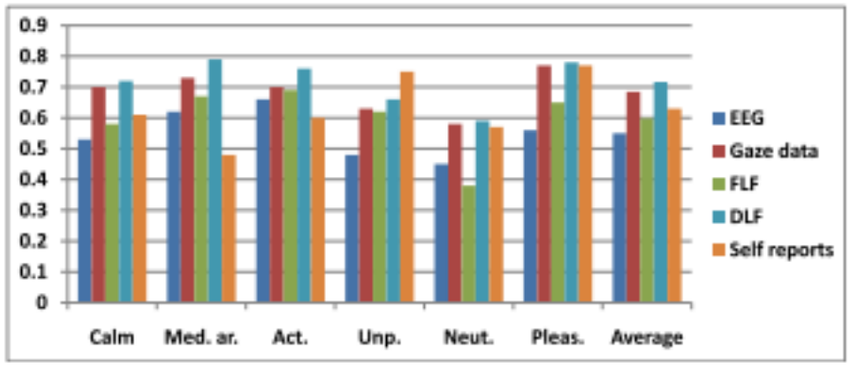


Fig. 9. This bar chart shows the F1 score for classification results of each class from different modalities.

Arousal classes were, on average, detected with higher accuracy using EEG signals comparing to valence labels. This might be due to higher visual and auditory variance of the arousal variant videos comparing to valence variant ones. Exciting scenes usually contain fast movements and loud noises, which manifest themselves both in EEG signals and pupillary responses, whereas the difference between pleasant and unpleasant responses can be hidden in the semantics. The direct bodily responses to different stimuli can increase the variance in responses and improve the

TABLE 5
Confusion Matrices of Different Classification Schemes (Row: Classified Label; Column: Ground Truth)

Arousal

	1	2	3
1	44	15	11
2	22	111	39
3	30	60	135

(a) EEG

	1	2	3
1	60	7	9
2	10	136	40
3	26	43	136

(b) Eye gaze data

	1	2	3
1	49	14	10
2	15	117	31
3	32	55	144

(c) FLF

	1	2	3
1	62	6	8
2	8	146	28
3	26	34	149

(d) DLF

	1	2	3
1	63	35	11
2	24	88	65
3	9	63	109

(e) Self reports

Valence

	1	2	3
1	87	56	52
2	23	52	15
3	54	31	97

(f) EEG

	1	2	3
1	108	46	26
2	36	77	12
3	20	16	126

(g) Eye gaze data

	1	2	3
1	139	83	56
2	7	36	10
3	18	20	98

(h) FLF

	1	2	3
1	115	48	22
2	27	75	12
3	22	16	130

(i) DLF

	1	2	3
1	126	40	4
2	38	91	53
3	0	8	107

(j) Self reports

The numbers on the first row and the first column of tables (a), (b), (c), (d), and (e) represent: 1) calm, 2) medium aroused, 3) activated and for tables (f), (g), (h), (i), and (j) represent: 1) unpleasant, 2) neutral valence, 3) pleasant. The confusion matrices relate to classification using (a), (f) EEG signals, (b), (g) eye gaze data, (c), (h) feature level fusion, (d), (i) decision level fusion, (e), (j) self-reports.

Open issue

- pupillary response for lights difference in age groups
- feature normalization
- single emotion values