Facial Expressions of Sentence Comprehension

Cigdem Turan[a], Yixin Wang[a], Shun-Cheung Lai[a], Karl David Neergaard[b], Kin-Man Lam[a]

[a] Dept. of Electronic and Information Engineering, [b] Dept. of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
Hong Kong SAR, China
{cigdem.turan, yixin.wang, shun-cheung.lai, karl.neergaard}@connect.polyu.hk, enkmlam@polyu.edu.hk

***Abstract*—Understanding facial expressions allows access to one's intentional and affective states. Using the findings in psychology and neuroscience, in which physical behaviors of the face are linked to emotional states, this paper aims to study sentence comprehension shown by facial expressions. In our experiments, participants took part in a roughly 30-minute computer mediated task, where they were asked to answer either "true" or "false" to knowledge-based questions, then immediately given feedback of "correct" or "incorrect". Their faces, which were recorded during the task using the Kinect v2 device, are later used to identify the level of comprehension shown by their expressions. To achieve this, the SVM and Random Forest classifiers with facial appearance information extracted using a spatiotemporal local descriptor, named LPQ-TOP, are employed. Results of online sentence comprehension show that facial dynamics are promising to help understand cognitive states of the mind.**

Keywords—sentence comprehension; affective computing; facial expression recognition; human-computer interaction

I. INTRODUCTION

With the great advancements in computer vision and machine learning techniques, a promising new literature is developing that uses dynamic facial expression data to interpret the facial expressions in the wild. The use of dynamic, multimodal data, while at its naissance, has a great deal yet to offer. Vural et al. [1], whose findings may lead to an early alarm system to avoid car accidents due to fatigue, studied the drowsiness expressions of drivers. Cohn et al. [2] analyzed the facial expressions of patients during psychoanalytic sessions with the intent of diagnosing depression. Recently, Dibeklioglu and Gevers investigated the automatic estimation of the level of taste liking through facial dynamics and showed that the proposed method is more reliable for estimating the taste than the human subjects [3]. Concurrent to studies in facial recognition, bodily movement is also being used to enrich dynamic and multimodal data. Jaques et al. [4] focused on understanding and predicting bonding between interlocutors during conversations using facial expressions and body language. A similar multimodal approach to recognition was used in videos of adults experiencing pain [5-8]. In 2017, Jaiswal et al. [9] detected the presence of ADHD/ASD using facial movements with the Kinect device, a tool that gives multiple streams of data for both facial and bodily movement.

An application, currently being explored with recognition technology, is facial expression in teaching environments [10-13]. The study in [12] proposed an affective e-learning model to investigate emotional states of learners and predict their future interaction with a learning system. Related affective states were based on a cognitive appraisal approach [14] that includes twenty emotions. Kort et al. [10] built a computer-based model that identifies users' affective states and responds accordingly, i.e. a learning companion. The paper also proposed a spiral affective model that combines the different phases of learning with the emotional axes. Closer to the goals of the current study, Sathik et al. [13] investigated student learning in a classroom setting through the correlation of successful comprehension with positive expressions, and failed comprehension with negative expressions. The authors did not build a model to predict students' comprehension but instead explored the statistical correlation of expressions towards comprehension types.

To date, none have investigated the multidimensionality of comprehension. Expanding the search to related fields brings us to research in psychology, where neuroscience tools matched with computational modelling, and experimentation, have laid a groundwork for the study of memory and learning specifically about language processing [15-18]. Facial behavior analysis has much to offer this growing literature in that it may well add another tool to the toolbox of identifying underlying states of cognition, much like the introduction of eye-tracking technology did over two decades ago [19-22]. The current study differs critically from previous work in that we analyze online aspects of comprehension as they occur dynamically across the face. We do not depend on labelled facial configurations of distinct emotions, nor positions within a continuous model, but rather analyze changes in facial configuration indicative of differing stages of comprehension. Analysis of comprehension according to individual facial dynamics can aid stakeholders (teachers, educators, etc.) in adapting educational practices through increasing access to learners' mental and affective states.

One area of investigation that allows for controlled yet natural responses to online stimuli is sentence processing. To fully comprehend a sentence, one must manage multiple different types of relationships, such as the morpho-syntactic and semantic-thematic relationships between the component parts of the sentence, and the relationship between the sentence's resulting meaning and its associations that have been encountered before and stored within long-term semantic memory [23]. Several studies in cognitive neuroscience have examined the neural networks that are active in processing these relationships by investigating the brain responses when each is violated, i.e. syntactic, semantic-thematic, and world knowledge [15-18, 23]. To study the facial expressions of

sentence comprehension, we choose only world knowledge violations in this study.

Experimental data was obtained in a behavioral experiment in which participants were asked to first read knowledge-based statements, second provide a 'true' or 'false' answer, and finally, receive feedback as either 'correct' or 'false'. During the experiment, the Kinect v2 device was used to record multiple streams of data. Using these data, we analyze the facial behavior not only during successful recall of target information, but during unsuccessful recall, as well as the expressions during guessing.

The organization of the rest of this paper is as follows. In Section II, the data collection method is explained in detail with respect to participants, stimuli and the experimental design. In Section III, the methodologies used in our experiments are explained. Section IV presents the experimental protocol and results. Finally, Section V concludes the paper.

II. DATA COLLECTION

In this section, the experimental design and the data collection process are explained in detail.

A. Participants

Forty-four healthy volunteers (twenty woman and twenty-four man) aged between 20 and 37 years (mean = 27, SD = 3.89) from 16 different nationalities participated in the study. All participants had normal or corrected-to-normal visual acuity. Self-rated English proficiency was also collected from the participants (mean = 7.61, SD = 1.57, with 10 as a native speaker). Participants gave informed written consent to the experimental procedure. This study was approved by the local ethics committee of The Hong Kong Polytechnic University.

B. Stimuli

The experimental material consisted of 100 true or false world-knowledge statements. Selection of the final stimuli began with 240 statements generated from a variety of general knowledge areas: mathematics, science, and technology, as well as art, entertainment, history and geography. Amazon Mechanical Turk was then used to assess each item's difficulty level based on accuracy rate. Four categories were created (25 Baseline TRUE, 25 Baseline FALSE, 25 Range TRUE, and 25 Range FALSE), in which Baseline statements were those that received a high accuracy rate, and Range statements were those that received roughly 50% accuracy, i.e., were the product of guessing. For the purposes of presentation, each statement was segmented into three parts. Table 1 presents example stimuli used in the experiment from each category.

TABLE I. EXAMPLE STATEMENTS ACCORDING TO THEIR TRUE/FALSE CATEGORY

Statements	Category
There are / 360 degrees / in a circle.	Baseline TRUE
An apple / is larger than / a grape.	Baseline TRUE
The Earth / rotates around / the Moon.	Baseline FALSE
Antarctica / is a province of / France.	Baseline FALSE
Dianne Wiest / won best supporting actress / in 1995.	Range TRUE
The Ig Nobel Prizes / have been awarded / since 1991.	Range TRUE
In 1912, / Jean Sibelius led / his Fourth Symphony premiere.	Range FALSE
Diet Coke / was invented / in 1970.	Range FALSE

C. Experimental Design and Procedure

The behavioural experiment was designed using PsychoPy, a stimulus delivery library for the Python programming language [24, 25]. PsychoPy allows for the online recording of participants' responses and response times. The experiment was designed as a self-paced reading task, in which participants control the progression of each experimental window, and therefore its duration, through the press of the space button. Each trial within the experiment consists of 6 windows, as can be seen in Fig. 1. In Window 0, the sentence stimuli are masked with dashes that are equal to the length of the sentence if displayed. In Window 1, only the first of three sentence segments is revealed, while the remaining segments remain masked. Windows 2 and 3 similarly display only the consecutive segments of the sentence while masking the remaining parts. In the Window Answer, participants are asked to evaluate the veracity of the statement by clicking one of the highlighted texts on the computer screen indicating "True" or "False". Finally, participants are given feedback in the Window Feedback as either "correct" or "incorrect". Each trial ends after a 500 millisecond inter-stimulus interval.

During the experiment, participants sat in front of a computer screen in a quiet room. A Kinect 2.0 device, capable of capturing high-resolution RGB videos, was placed on top of the computer screen to save the relevant information during the experiment. All animation units, as well as RGB, depth, and infrared streams obtained from the Kinect 2.0 device,

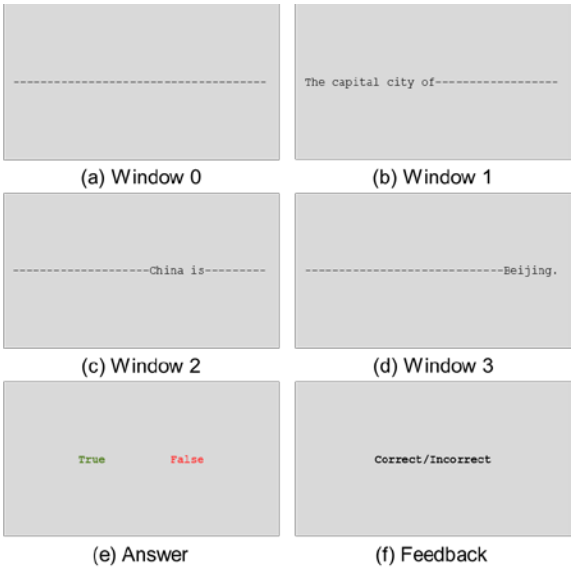

Figure 1. Experimental procedure

TABLE II. PARTICIPANTS' AVERAGE ACCURACIES AND STDS

Questions	Mean Acc.	Standard dev.
All	0.7023	0.0630
Baseline	0.8900	0.0845
Range	0.5145	0.0751

were saved to a hard drive during the experiment.

III. METHODOLOGY

At the cognitive level, the construction of an interpretation of a sentence requires not only combining the meaning of words and phrases followed by computing their thematic and syntactic relations, but also using world knowledge [26]. At the brain level, sentence comprehension activates a network of neurons whose activation areas and degrees differ with the type and the complexity of the sentence. The psycholinguists and neuroscientist have been extensively investigated sentence comprehension in the past with EEG and fMRI [15, 23, 27]. Recently, the studies have been started using eye tracking technology [19-22]. This study aims to introduce a new tool to understand the sentence comprehension through facial dynamics and vice versa.

In the first part of the behavioural experiment design, we manipulated comprehension by altering the veracity of the statements to investigate the facial dynamics during online sentence comprehension, in case of world-knowledge violations. We also controlled the difficulty of the world-knowledge to explore the facial dynamics when the participants lack the required knowledge. It impels us to analyse participants' facial behaviours during online sentence comprehension in terms of two aspects: 1) whether we can identify that they knew the question of interest or simply guessed, i.e. knowing face versus guessing face, and 2) whether we can identify the veracity of the statements through their facial dynamics.

To achieve the analysis, the faces in the videos are aligned based on eye points and cropped to 180×180 pixels. Then, local features from the eye and the mouth regions are extracted separately by a spatiotemporal descriptor, called "Local Phase Quantization Three Orthogonal Planes" (LPQ-TOP) [28], with the Support Vector Machine (SVM) classifier and Random Forest (RF) classifier.

LPQ-TOP, which is a spatiotemporal extension of Local Phase Quantization (LPQ) into three orthogonal planes (TOP), was selected, because we found that LPQ is an effective and discriminative feature of facial expression and has achieved excellent performance on facial expression recognition [29]. Facial expression for comprehension can be viewed as an expression in between micro and macro-expressions. TOP is an effective feature-extraction method for micro-expression recognition [30].

To analyse the effect of each window, the spatiotemporal local features are extracted from frames belonging different window configurations, e.g. w23 as Windows 2 and 3 and wA as Window Answer, by discarding Window 1 because of two reasons: 1) the residual of the facial configuration caused by the feedback often observed in the beginning of the next trial, and 2) the part of the sentence given in Window 1 does not affect the truth condition of the statement.

IV. EXPERIMENTAL PROTOCOL AND RESULTS

A. Experimental Protocol

The data collected from the Kinect 2.0 device is divided into segments, where each segment has the Animation Units and the frames belonging to only one trial answered by one participant, i.e. per trial and per participant. It means that one participant would have maximum 100 RGB video shots, 100 depth video shots, etc. corresponding to 100 stimuli.

We cleaned up the database by deleting segments where the participants got distracted by external factors or the RGB videos weren't saved correctly. Therefore, not all the participants have 100 unique segments, although their responses to all of the statements were saved using the PsychoPy experiment. Also, we discarded one of the participants since the participant does not have enough segments to run our experiments. As a result, the experimental data consists of 3,257 unique segments from 43 participants and 100 stimuli where each unique segment has the RGB, depth and infrared video shots, as well as 17 animation units per frame provided by the Kinect 2.0 device. The average length of the videos is 9.1990 seconds (STD: 3.6962 seconds).

In the first set of experiments, we assume that all the trials belonging to the Baseline class, and answered correctly by the participants, represent the class "knowing face". Similarly, all the trials belonging to the Range class should be part of the class "guessing face". The stimuli representing false world-knowledge in the Baseline are also discarded to prevent the facial dynamic caused by the veracity of the statements.

In the second set of experiments, all trials belonging to the Baseline class and answered correctly by the participants are further divided into two classes: representing the veracity of the statements as "true" and "false". Here, we aim to detect the

TABLE III. PARTICIPANTS' AVERAGE ACCURACIES

AUC (%)		EYE				MOUTH				EYE & MOUTH			
		w23	**w2**	**w3**	**wA**	**w23**	**w2**	**w3**	**wA**	**w23**	**w2**	**w3**	**wA**
Knowing face vs. Guessing face	**Random Forest**	57.54	57.05	54.62	55.36	55.48	55.92	54.31	**57.54**	56.67	56.32	54.84	**56.70**
	SVM	22.05	24.54	29.68	**32.60**	16.82	18.69	26.43	**28.16**	28.09	28.18	32.98	**37.91**
True vs. False	**Random Forest**	46.55	48.99	45.70	**78.92**	47.92	47.11	45.75	**67.77**	**49.23**	46.31	46.78	46.57
	SVM	**54.87**	51.56	50.13	50.20	**56.17**	54.36	51.07	52.64	**50.57**	48.08	44.52	47.95

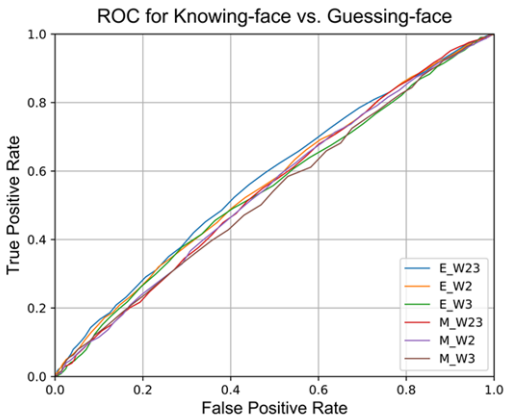

Figure 2. The ROC of the experimental results obtained using the Random Forest classifier on the detection of the knowing face

statements with world-knowledge violations using facial dynamics.

Because of the fact that the participants in our experiment were from 17 different countries and with different cultural backgrounds and languages, so the facial dynamics in our study are not generalizable across the participants. Therefore, the leave-one-trial-out (LOTO) classification scheme is adopted, per person, to investigate the predictability of a knowing face and the veracity of the statements. The LOTO scheme trains each fold using the $n_s - 1$ trials belonging to one person and tests the model with one trial, where n_s is the number of trials belonging to the s-th participant.

We use two different classification methods to investigate the recognition rates of the two sets of experiments: the SVM classifier and the RF classifier. Since the number of samples per subject is much less than the dimensionality of the features, i.e. the dimensionality of the LPQ-TOP features is 4,608 in our experiments, a subspace-learning method, named Marginal Fisher Analysis (MFA) [31], is applied for dimensionality reduction. MFA, which is a genral framework for manifold learning and dimensionality reduction, constructs two adjacency graphs to represent the within-class and the between-class geometry of the data and uses the Fisher criterion.

It is worth noting that MFA is applied only to the experiments that used the SVM classifier. The RF classifier can deal with high dimensionality with an increased number of trees.

B. Experimental Results

Each participant was confronted with 100 true/false statements. Table II shows their average accuracies and standard deviations with respect to all statements, Baseline and Range. As tested and observed in Amazon Mechanical Turk, Baseline statements achieved higher accuracies and Range statements were answered with a correction rate of 51% only by the participants of the experiment. One reason for Baseline statements not reaching one hundred percent accuracy, unlike our first goal, is the participants' English proficiency, since not all of the participants were native English speakers – the mean of self-rated English proficiency being 7.61.

Table III shows the AUCs, i.e. areas under the ROC, of the first and second set of experiments on online sentence comprehension, where Figure 2 presents the ROC of the experimental results using the RF classifier on the detection of knowing faces. As observed in Table III and Figure 2, the highest accuracy is obtained by using features from the eye and the mouth windows during Windows 2 and 3.

We observed in Table III that the best result on knowing face vs. guessing face is achieved by using the mouth features during the Window Answer (wA), with the RF classifier. For true vs. false, the best result is still obtained when wA and the RF classifier are used, but the eye window should be considered. This suggests that the expression of a knowing face and that of the world-knowledge violation in a statement are reflected in different facial features.

Experiment results also show that the SVM classifier fails to detect the knowing face since it suffers from overfitting, although a supervised subspace-learning method is applied to avoid overfitting.

V. CONCLUSION

Researchers in different disciplines, such as computational linguistics and computational neuroscience, are often not aware of the advances in recognizing facial information. This disconnection between disciplines limits the interdisciplinary multimodal studies to understand human facial behavior. This study motivates and fosters the interdisciplinary study by attempting to bring together studies, results and questions from different disciplines by focusing on the computational analysis of human behavior in an experimental setting, specifically facial behavior which can practically provide methodological support to investigate people's facial behavior and mental states.

We collected 100 general knowledge questions from a wide variety of topics, including mathematics, history, sports, art, etc. A total of 44 participants joined our behavioral experiment, where they were asked to answer the collected questions as true or false in front of a computer, while the Kinect v2 device was recording their faces. After each response, the participants were shown their achievements as correct or incorrect.

Using the videos obtained during the behavioral experiment, we conducted experiments using the SVM and the Random Forest classifiers to investigate the online sentence comprehension through facial expression in two stages. The results show that the eye and mouth windows, during the answering window, can best be used to detect the mental states of online sentence comprehension.

ACKNOWLEDGMENT

The authors wish to thank Stephen Politzer-Ahles for his insightful critics and suggestions on the design of the behavioral experiment. The work described in this paper was supported by a research grant from The Hong Kong Polytechnic University (project code: G-YBKF).

REFERENCES

[1] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan, "Drowsy driver detection through facial movement analysis," Human–computer interaction, pp. 6-18, 2007.

[2] J. F. Cohn et al., "Detecting depression from facial actions and vocal prosody," in 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009), 2009, pp. 1-7: IEEE.

[3] H. Dibeklioglu and T. Gevers, "Automatic Estimation of Taste Liking through Facial Expression Dynamics," IEEE Transactions on Affective Computing, 2018.

[4] N. Jaques, D. McDuff, Y. L. Kim, and R. Picard, "Understanding and Predicting Bonding in Conversations Using Thin Slices of Facial Expressions and Body Language," in International Conference on Intelligent Virtual Agents, 2016, pp. 64-74: Springer.

[5] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," Advances in visual computing, pp. 368-377, 2012.

[6] A. B. Ashraf et al., "The painful face–pain expression recognition using active appearance models," Image and vision computing, vol. 27, no. 12, pp. 1788-1796, 2009.

[7] P. Lucey et al., "Automatically detecting pain in video through facial action units," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 41, no. 3, pp. 664-674, 2011.

[8] P. Rodriguez et al., "Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification," IEEE Transactions on Cybernetics, 2017.

[9] S. Jaiswal, M. F. Valstar, A. Gillott, and D. Daley, "Automatic detection of ADHD and ASD from expressive behaviour in RGBD data," in 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 2017, pp. 762-769: IEEE.

[10] B. Kort, R. Reilly, and R. W. Picard, "An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion," in Proceedings of IEEE International Conference on Advanced Learning Technologies, 2001, pp. 43-46: IEEE.

[11] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser, "Facial features for affective state detection in learning environments," in Proceedings of the Cognitive Science Society, 2007, vol. 29, no. 29.

[12] L. Shen, M. Wang, and R. Shen, "Affective e-learning: Using" emotional" data to improve learning in pervasive learning environment," Journal of Educational Technology & Society, vol. 12, no. 2, p. 176, 2009.

[13] M. Sathik and S. G. Jonathan, "Effect of facial expressions on student's comprehension recognition in virtual educational environments," SpringerPlus, vol. 2, no. 1, p. 455, 2013.

[14] A. Ortony, G. L. Clore, and A. Collins, The cognitive structure of emotions. Cambridge university press, 1990.

[15] P. Hagoort, L. Hald, M. Bastiaansen, and K. M. Petersson, "Integration of word meaning and world knowledge in language comprehension," Science, vol. 304, no. 5669, pp. 438-441, 2004.

[16] S. Harris, S. A. Sheth, and M. S. Cohen, "Functional neuroimaging of belief, disbelief, and uncertainty," Annals of neurology, vol. 63, no. 2, pp. 141-147, 2008.

[17] J. F. Marques, N. Canessa, and S. Cappa, "Neural differences in the processing of true and false sentences: Insights into the nature of 'truth'in language comprehension," Cortex, vol. 45, no. 6, pp. 759-768, 2009.

[18] W. Groen et al., "Semantic, factual, and social language comprehension in adolescents with autism: an FMRI study," Cerebral Cortex, vol. 20, no. 8, pp. 1937-1945, 2009.

[19] V. Demberg and F. Keller, "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity," Cognition, vol. 109, no. 2, pp. 193-210, 2008.

[20] T. Armstrong and B. O. Olatunji, "Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis," Clinical psychology review, vol. 32, no. 8, pp. 704-723, 2012.

[21] R. Jacob and K. S. Karn, "Eye tracking in human-computer interaction and usability research: Ready to deliver the promises," Mind, vol. 2, no. 3, p. 4, 2003.

[22] T. Van Gog and K. Scheiter, "Eye tracking as a tool to study and enhance multimedia learning," ed: Elsevier, 2010.

[23] G. R. Kuperberg, T. Sitnikova, and B. M. Lakshmanan, "Neuroanatomical distinctions within the semantic system during sentence comprehension: evidence from functional magnetic resonance imaging," Neuroimage, vol. 40, no. 1, pp. 367-388, 2008.

[24] J. W. Peirce, "PsychoPy—psychophysics software in Python," Journal of neuroscience methods, vol. 162, no. 1, pp. 8-13, 2007.

[25] J. W. Peirce, "Generating stimuli for neuroscience using PsychoPy," Frontiers in neuroinformatics, vol. 2, p. 10, 2009.

[26] M. A. Just, P. A. Carpenter, T. A. Keller, W. F. Eddy, and K. R. Thulborn, "Brain activation modulated by sentence comprehension," Science, vol. 274, no. 5284, pp. 114-116, 1996.

[27] S. M. Garnsey, M. K. Tanenhaus, and R. M. Chapman, "Evoked potentials and the study of sentence comprehension," Journal of psycholinguistic research, vol. 18, no. 1, pp. 51-60, 1989.

[28] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011, pp. 314-321: IEEE.

[29] Cigdem Turan and Kin-Man Lam, "Histogram-based local descriptors for facial expression recognition (FER): A comprehensive study," Journal of Visual Communication and Image Representation, 2018.

[30] Guoying Zhao and Matti Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 915-928, June 2007.

[31] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," IEEE Trans on Pattern Analysis and Machine Intelligence, vol. 29, no. 1, pp. 40-51, 2007.