Personality Trait Classification Based on Co-occurrence Pattern Modeling with Convolutional Neural Network

Ryo Kimura and Shogo Okada(✉)

Japan Advanced Institute of Science and Technology, Nomi, Japan
okada-s@jaist.ac.jp

Abstract. In the modeling of impressions, a key factor for success is to extract nonverbal features that can be used to infer the target variable. To extract the effective features for capturing the relationship between the target subject which has the personality trait and other group members, Okada et al. propose a co-occurrence event-mining framework to explicitly extract the inter-modal and inter-personal features from multimodal interaction data. The framework is an unsupervised feature extraction algorithm by considering the relationship between nonverbal patterns of the target subject and group members. Though the label data of personality trait is useful to improve the accuracy, the valuable label data is not used for feature extraction. In this paper, we enhance the inter-modal and inter person feature extraction algorithm by using a deep neural network. We proposed a representation learning algorithm for capturing inter-modal and inter-person relationships by integrating using a convolutional neural network (CNN). In the experiment, we evaluate the effectiveness of the representation learning approach using the ELEA (Emerging Leadership Analysis) corpus, which includes 27 group interactions and is publicly available. We show that the proposed algorithm with CNN slightly improves the personality trait classification accuracy of the previous algorithms. In addition, we analyze which slice of multimodal time-series data is key descriptors to predict the personality trait using the proposed algorithm with CNN.

Keywords: Impression · Multimodal interaction · Personality traits · Convolutional neural network

1 Introduction

In recent comprehensive research on the computational multimodal analysis, the modeling of impressions is the focus of attention. The target variables vary widely, such as public speaking skills [1,2], persuasiveness [3], communication skill in job interviews [4], and leadership [5]. A key factor for success is to extract nonverbal features that can be used to infer the target variable. To extract the effective features, previous works have defined static features from audio and

C. Stephanidis et al. (Eds.): HCII 2020, LNCS 12427, pp. 359–370, 2020.
https://doi.org/10.1007/978-3-030-60152-2_27

visual data based on knowledge of social science. Speaking activity and prosodic features as audio cues, body activity, head activity, hand activity, gaze and facial expression as visual cues, are used for inference of personality traits. However such statistic features ignore the dynamics of nonverbal events observed in the whole meeting and the relationship between the target subject which has the personality trait and other group members. To solve the problem, Okada et al. [6,7] propose a co-occurrence event-mining framework to explicitly extract the inter-modal and inter-personal features from multimodal interaction data. The key approach of the framework is to discover co-occurrence patterns between modalities. The accuracy of the model trained with the inter- modal and inter person features outperforms that of models trained with the traditional statistic feature set. From another viewpoint, the framework is an unsupervised feature extraction algorithm by considering the relationship between nonverbal patterns of the target subject and group members. Though the label data of personality trait is useful to improve the accuracy, the valuable label data is not used for feature extraction. In this paper, we enhance the inter-modal and inter person feature extraction algorithm by using a deep neural network. We proposed a representation learning algorithm for capturing inter-modal and inter-person relationships by integrating using a convolutional neural network (CNN). In the experiment, we evaluate the effectiveness of the representation learning approach using the ELEA (Emerging Leadership Analysis) corpus, which includes 27 group interactions and publicity available. We show that the proposed algorithm with CNN improves the personality trait classification accuracy of the previous algorithm proposed in [6,7].

2 Related Work

Our research is related to personality-trait modeling and interaction mining. This study focuses on impression modeling in conversations. For multiparty interactions, different works included different variables: social roles [8,9], dominance [10], personality traits [11,12] and leadership [5]. As a common approach of these works, audio, and visual features are calculated using the mean, median, min, max, and X percentile of various statistics (count and length) from each pattern observed throughout an entire meeting or for a part of a meeting [5,11,13]. Although this approach can often fuse the total statistics of patterns observed within a specified duration, it cannot capture co-occurrence between multimodal patterns for each time period. For example, extracting co-occurrence events between an utterance and a body-motion pattern as a feature is useful if the utterance accompanying the body gesture makes a stronger impression on the listener than that utterance without the gesture. Our mining algorithm explicitly extracts such co-occurrence features. Several other studies have focused on extracting the correlations between modalities. Song et al. [14] proposed a multimodal technique that models explicit correlations among modalities via canonical correlation analyses (CCAs) [15]. The algorithm was evaluated using a recognition task for disagreement/agreement with a speaker in political debates

[16]. Chatterjee et al. [17] proposed an ensemble approach that combines a classifier based on inter-modality conditional independence with a classifier based on dimension reduction via a multiview CCA. Feature co-occurrence is often adopted in computer vision [18–23].

Preliminary works [6,24,25] have been performed using co-occurrence pattern mining similar to the proposed approach. Okada et al. [24] used a co-occurrence pattern-mining algorithm, which is a modified version of the algorithm in [26], to extract features to infer the performance level of storytelling in group interaction. The main difference with respect to our work is that the research focuses on the modeling of group performance and not the individual performance and that nonverbal features are extracted manually. The main limitation of these research works [24,25] is that only binary event (on/off) features are used for mining. Okada et al. [7] enhanced the co-occurrence pattern mining algorithm and also applied the algorithm for dyadic-interaction dataset. The enhanced algorithm proposed in [7] improves the classification accuracy of personality traits. In this paper, we enhance the inter-modal and inter person feature extraction algorithm by using a deep neural network. We proposed a representation learning algorithm for capturing inter-modal and inter-person relationships by integrating using a convolutional neural network (CNN). The main contribution of this paper is to analyze when the effective co-occurrence features are observed in the whole meeting using the co-occurrence pattern learning algorithm with CNN.

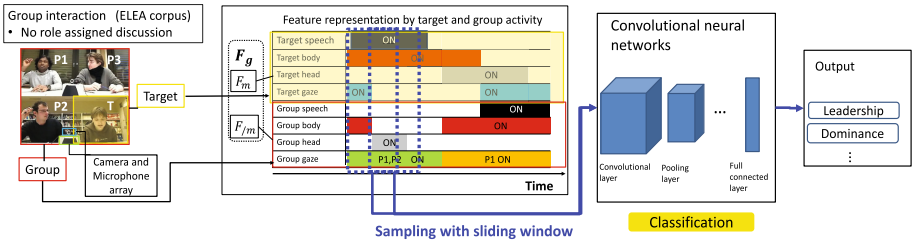

Fig. 1. Overview of proposed framework

3 Inter-person and Inter-modal Representation Learning

Figure 1 shows an overview of the proposed framework. We proposed Inter-person and Inter-modal Representation Learning by using convolutional neural networks, which are mainly used in the field of image recognition. The convolutional neural network extracts feature with a sliding window method applied to time-series data. Therefore, convolutional neural networks are expected to capture time-series dependency than feature extraction by co-occurrence patterns. In addition, there is a possibility that data features can be accurately captured by performing supervised learning using label data in feature extraction. However, the input of the convolutional neural network does not support multimodal

features. Therefore, it is necessary to convert data with multimodal features into a form that can be used as input.

3.1 Multimodal Feature Representation

We propose a feature representation method for capturing the co-occurrence of the nonverbal patterns observed for each participant. We define co-occurrence patterns as multimodal events that overlap in time. Each event has a time length and corresponds to a segment denoted by "ON" in Fig. 1. We define an event as a segment in which the feature is active. Multimodal features are represented as follows. The feature representation for group interaction is described. We propose a feature representation for comparing nonverbal patterns that are observed for each participant in a group. The representation captures how a participant acts when other members execute any nonverbal activity by simultaneously observing the nonverbal activities of both the individual participant and the other group members. Let F_{group} be the feature set for a group interaction:

$$\boldsymbol{F}_{group} = \{F_m, F_g\}. \tag{1}$$

F_m is the feature representation for one specific person in a group, and F_g is the feature representation for a group composed of the other members without m. An example of $\{F_m,\ F_g\}$ is shown in Fig. 1. The co-occurrence pattern mining requires conversion of the time-series signal data into a sequence of events $(f_{m,i})$ with a finite time length as a preprocessing step. Multimodal behavior is inherently observed as time-series signals in a session. The binarization or discretization of continuous time-series data is described in Sect. 5. The modified audio-visual features f in F_{group} are also described at the bottom of Fig. 2, respectively.

3.2 Convolutional Neural Network (CNN) to Learn Co-Occurrence Pattern

The data type of multimodal features extracted from group meetings is multidimensional time series, so data is represented as two-dimensional matrix data with the vertical axis representing the number of features and the horizontal axis representing time. We segment the time-series data into slices with almost 1 min (58 s) and the two-dimensional data with $D \times 58$ s is defined as a training or test sample to input into CNN. The converted data is binary data in which the time-segment where the feature is recognized as "1" and the segment where the feature is not recognized is "0". The method for feature extraction is described in Sect. 5. The converted data is used as input to the convolutional neural network. The label data includes five labels for leadership ability evaluated on a seven-point scale. In this paper, we perform experiments of the classification task by replacing this label data with binary data that is above or below the average of participants.

4 Dataset and Features

4.1 ELEA: Group-Interaction Dataset

We used a subset of the ELEA corpus [5] for this study. The subset consists of audio-visual (AV) recordings of 27 meetings in which the participants performed a winter survival task with no roles assigned. A total of 102 participants were included (six meetings with three participants and 21 meetings with four participants). Each meeting lasted approximately 15 min. The synchronization of audio and video was performed manually by aligning the streams according to the clapping activity. Additional details on the ELEA AV corpus can be found in [27].

ID	Features	Symbol	Description
F_1	Speaking Status (ST)	ST	Speech segments of the target person
		$SO1$	One person other than target speaks
		$SO2$	More than two people speak.
		$Ssil$	Silent segment
F_2	Pitch (PI)	PUp, PDo	Sign of difference between utterance t and utterance $t-1$
		PCL, PCM, PCH	Cluster index (low medium and high level) after clustering
		$PCNL, PCNM, PCNH$	Cluster index after clustering of normalized value
F_3	Energy (EN)	EUp, EDo	Sign of difference between utterance t and utterance
		ECL, ECM, ECH	Cluster index after clustering
		$ECNL, ECNM, ECNH$	Cluster index after clustering of normalized value
F_4	Head Motion (H)	HMT	Motion segments of target person
		$HMO1$	One person other than target moves
		$HMO2$	More than two people move
		$HMsil$	Still motion segment
F_5	Body Motion (B)	BMT	Motion segments of target person
		$BMO1$	One person other than target moves
		$BMO2$	More than two people move
		$BMsil$	Still motion segment
F_6	MEI (MT)	MUp, MDo	Sign of difference between segments
		MCL, MCM, MCH	Cluster index after clustering
		$MCNL, MCNM, MCNH$	Cluster index after clustering of normalized MEI
F_7	Gaze (G)	GT	Target person looks at person
		$GTSp$	Target person looks at speaker
		$GOT1$	One person looks at the target
		$GOT2$	More than two people look at the target
		MGT	Mutual gaze between target and another person
		MGO	Mutual gaze between two people other than target

Fig. 2. Multimodal feature set [6] (The feature set used in this study is aligned to that used in [6] for comparing the accuracy of the proposed framework with that of [6]. This table is adopted by the article [6]).

The ELEA corpus also includes scores for traits of individuals with respect to dominance and leadership. After the meeting task, the participants completed a Perceived Interaction Score, which captures perceptions from participants during the interaction, in which they scored every participant in the group based

on four items related to the following concepts: "Perceived Leadership (Leadership)", "Perceived Dominance (Dominance)", "Perceived Competence (Competence)" and "Perceived Liking (Liking)". Afterward, the "Dominance Ranking (Ranked Dominance)". Leadership captures whether a person directs the group and imposes his or her opinion. Dominance captures whether a person dominates or is in a position of power. Participants were asked to rank the group, assigning 1 to the most dominant participant and 3 or 4 to the less dominant participants. Additional details can be found in [5].

5 Multimodal Features

Multimodal features are extracted automatically from audio and visual cues in this study in same manner with [6,11]. The feature sets of the ELEA used in this study are summarized in Fig. 2. The detail of multimodal features used in this study is described in [6,7,11].

5.1 Audio Features

Speaking Status. Binary segmentation is performed to capture the speaking status (ST) of each participant. This binary segmentation is provided by the microphone array, and all speaking activity cues are based on the speaker segmentations obtained using the Microcone, which is used for the audio recordings and speaker diarization in [5,11] and [4]. We define a set of segments in which the speech status is "on" as the speaking-turn set ST.

Prosodic Features. Prosodic features are extracted for each individual member. Based on the binary speaker segmentation, we obtain the speech signal for each participant. Overlapping speech segments are discarded, only the segments in which the participant is the sole speaker are considered for further processing. Three prosodic speech features (energy, pitch) are determined based on the signal. We calculate the sign of the difference between the statistics of utterance j and utterance $j + 1$ using statistical t-test. Energy and pitch signals are converted into three categorical data (low level, middle level and high level) using k-means clustering.

5.2 Visual Features

Visual Activity Features. The first approach is based on head and body tracking and optical flow, which provides the binary head and body activity status and the amount of activity as well. As done for speech states, binary segmentation is done and an activity state set is extracted for head and body motion.

Motion Template Based Features. As a second approach, we have used Motion Energy Images (MEI) [28] as descriptors of body activity. We used the length of the meeting segment to normalize the images. Motion Energy Images (MEI) are obtained by integrating each difference image from whole video clip. Significant changes of MEI have the possibility to capture behaviors related to personality traits. The features are extracted in same manner with prosodic features

Visual Focus of Attention Features. Visual focus of attention (VFOA) features were extracted and shared by the authors in [27], where a probabilistic framework was used to estimate the head location and pose jointly based on a state space formulation. We define a set of segments GT where the target participant looks at the other participants through the meeting. We also define a set of segments $GTSp$ where the target participant looks at the speaker. Looking at speaker is an important signal of the listener's interest and politeness [29]. We further define two features $GOT1, GOT2$ as group attention features $G_{/m}$. $GOT1$ is a set of segments where one member looks at the member m. $GOT2$ is a set of segments where more than two members looks at the member m.

Next, we extracted mutual gazing features (although mutual gazing is defined as co-occurrence pattern with GT and $GOT1, 2$). We prepare two group features for mutual gazing. MGT is a set of segments where one member x looks at the member m and vice versa. MGO is a set of segments where two members y, z except the member m look at each other.

6 Experiments

6.1 Experimental Setting

For the ELEA group interaction, the inference tasks were classification and regression in [11], and then further studied as classification in [6]. In this paper, we decided to focus only on the binary classification task in the same manner with [6,7], because the objective of this experiment is to compare with algorithms proposed in [6,7,11]. We classify binary levels of the impression index. In the classification task, impression values are converted to binary values (high or low) by thresholding using the median value. For example, this method is performed to represent people scoring high/low in terms of leadership. The trained model is evaluated based on the classification accuracy of the test data. In the experiments presented below, we use leave-one-out cross-validation and report the average accuracy over all folds. We normalize the data such that each feature has a zero mean and one standard deviation.

6.2 Setting of Proposed Algorithm

The network structure of CNN is shown in Fig. 3. Rectified linear function (Relu) is used as the activation function in all middle layers and cross-entropy function is

used as the loss function. The loss function is defined as $E = -\frac{1}{N}\Sigma q(k)log(p(k))$, where $p(k)$ denote probability of each label for sample x_k, which is output by the CNN and $q(k)$ denotes the ground-truth distribution for sample x_k. In the testing phase, the output probability per time-series slice of multimodal features is output from CNN. The classification result for the slice is correct when the class with the highest output probability is equal to the true label. N time-series slices are obtained from a meeting (or a participant) and the classification accuracy is calculated as $\frac{Num.\,of\,correctly\,classified\,samples}{N}$ for each participant. If the classification accuracy is more than 0.5, we define that the classification for a participant is correct.

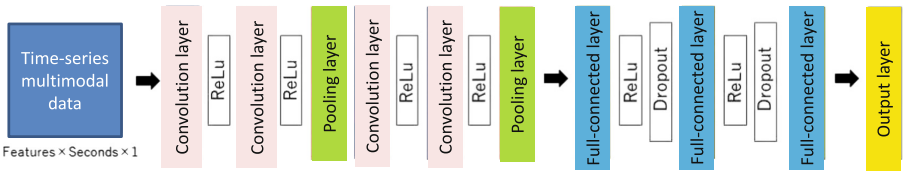

Fig. 3. Network structure of CNN

Table 1. Classification accuracy for leadership traits

	Perceived Leadership	Perceived Dominance	Perceived Competence	Perceived Liking	Ranked Dominance
Best in [7]	**73.53**	55.88	56.86	65.69	61.76
Best of [6]	72.55	61.76	**64.71**	53.92	64.71
Best of [11]	72.55	65.69	52.94	64.71	51.96
Co-occur CNN (All)	70.56	55.88	52.94	61.76	**66.67**
Co-occur CNN (Target)	64.71	**66.67**	50.98	**66.67**	**66.67**

6.3 Experimental Results

The Table 1 shows the classification accuracy for 5 leadership traits in ELEA corpus. We compared the accuracy of proposed models: Co-occur CNN (All) and Co-occur CNN (Speaker) with the best accuracy of proposed models of [6,7,11], which are reported in these articles. Co-occur CNN (All) is the CNN trained from time-series data which is composed of multimodal features observed from all members in a group. Co-occur CNN (Speaker) is the CNN trained from time-series data which is observed from only the target person who is the subject for the trait classification. From the Table 1, Co-occur CNN (Target) which is a proposed method obtained the best accuracy for "Perceived Dominance", "Perceived Liking" and "Ranked Dominance" with 66.67%. The proposed method improved the accuracy with 1–2 point. On the other hands, Best accuracy for

"Perceived Leadership" is obtained by [7] with 73.53% and that for "Perceived Competence"is obtained by [6] with 64.71%. These results show that applying CNN for co-occurrence pattern modeling is effective to improve the accuracy, though the improvement is limited. The reason why the improvement is limited is discussed as follows. The proposed framework regards time-series slices as independent training samples for input to CNN. Though some samples (time-series slices) are useful for improving the accuracy because the multimodal features in the slice can capture the personality traits of participants. Contraversely, samples are unnecessary as training samples if the observed slices (multimodal features) are noise data.

6.4 Time-Series Analysis of Classification Accuracy

The proposed algorithm with CNN enables us to analyze which slices of multimodal time-series data are key descriptors to predict the personality trait. In the proposed method, multimodal time-series slice is input to CNN and the slice is classified into binary classes. We can analyze when the key multimodal features are observed while the group meeting by comparing the classification accuracy per the time-series slice. Table 2 shows the classification accuracy per time-series slice. The horizontal axis denotes time-series and the vertical axis denotes the type of personality traits. The accuracy for the time segment ("T(X)" in Table 2) denotes the mean accuracy which is calculated by averaging accuracy over six slices (almost 6 min). The bold values denote the best and second-best accuracy for each trait. For "Perceived Leadership" and "Perceived Liking", the accuracy of T 4–6 is better than that of other segments. These results mean that the effective multimodal features are observed in the middle of the meeting, so the accuracy of a middle zone (T 4–6) tends to be higher than others. The accuracy of T 1 and T9 (62%, 61%) in "Ranked Dominance" and accuracy of T 8, 9 (59%, 55%) in "Perceived Competence" is better than that of other segments. These results mean that effective multimodal features are observed at the start of a meeting or at the end of the meeting. The effective features are observed in different timing per the types of traits.

Table 2. Classification accuracy per time-series slices (The accuracy for time segment ("T(X)" in Table 2) denotes the mean accuracy which is calculated by averaging accuracy over six slices (almost 6 min). The bold values denote the best and second-best accuracy for each trait.

Co-occur CNN (All)	T(1)	T(2)	T(3)	T(4)	T(5)	T(6)	T(7)	T(8)	T(9)
Leadership	66	65	65	**70**	**69**	65	64	59	59
Dominance	52	**57**	**57**	54	56	**57**	56	55	56
Competence	50	51	48	50	49	54	49	**59**	**55**
Liking	55	56	54	**57**	54	**60**	54	55	52
Ranked Dominance	**61**	60	59	60	60	54	59	56	**62**

7 Conclusion

In this paper, we enhance the inter-modal and inter person feature extraction algorithm by using a deep neural network. We proposed a representation learning algorithm for capturing inter-modal and inter-person relationships by integrating using a convolutional neural network (CNN). In the experiment, we evaluate the effectiveness of the representation learning approach using the ELEA (Emerging Leadership Analysis) corpus. The experimental results show the classification accuracy for 5 leadership traits in ELEA corpus. We compared the accuracy of proposed models: Co-occur CNN (All) and Co-occur CNN (Speaker) with the best accuracy of proposed models of [6,7,11], which are reported in these articles. Co-occur CNN (Target) which is a proposed method obtained the best accuracy for "Perceived Dominance", "Perceived Liking" and "Ranked Dominance" with 66.67%. The proposed method improved the accuracy with 1–2 point. Through time-series analysis, we found that the effective features are observed in different timing per the types of traits. The future work is to improve the accuracy by finding the effective features with using an attention mechanism.

References

1. Wörtwein, T., Chollet, M., Schauerte, B., Morency, L.P., Stiefelhagen, R., Scherer, S.: Multimodal public speaking performance assessment. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), New York, NY, USA, pp. 43–50 (2015)
2. Ramanarayanan, V., Leong, C.W., Chen, L., Feng, G., Suendermann-Oeft, D.: Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 23–30 (2015)
3. Park, S., Shim, H.S., Chatterjee, M., Sagae, K., Morency, L.P.: Computational analysis of persuasiveness in social multimedia: a novel dataset and multimodal prediction approach. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), New York, NY, USA, pp. 50–57 (2014)
4. Nguyen, L., Frauendorfer, D., Mast, M., Gatica-Perez, D.: Hire me: computational inference of hirability in employment interviews based on nonverbal behavior. IEEE Trans. Multimed. **16**(4), 1018–1031 (2014)
5. Sanchez-Cortes, D., Aran, O., Mast, M.S., Gatica-Perez, D.: A nonverbal behavior approach to identify emergent leaders in small groups. IEEE Trans. Multimed. **14**(3), 816–832 (2012)
6. Okada, S., Aran, O., Gatica-Perez, D.: Personality trait classification via co-occurrent multiparty multimodal event discovery. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 15–22 (2015)
7. Okada, S., Nguyen, L.S., Aran, O., Gatica-Perez, D.: Modeling dyadic and group impressions with intermodal and interperson features. ACM Trans. Multimed. Comput. Commun. Appl. **15**(1s), 1–30 (2019)
8. Vinciarelli, A.: Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. IEEE Trans. Multimed. **9**(6), 1215–1226 (2007)

9. Zancanaro, M., Lepri, B., Pianesi, F.: Automatic detection of group functional roles in face to face interactions. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 28–34 (2006)
10. Rienks, R., Heylen, D.: Dominance detection in meetings using easily obtainable features. In: Proceedings of the International Workshop on Machine Learning for Multimodal Interaction, pp. 76–86 (2005)
11. Aran, O., Gatica-Perez, D.: One of a kind: inferring personality impressions in meetings. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 11–18 (2013)
12. Pianesi, F., Mana, N., Cappelletti, A., Lepri, B., Zancanaro, M.: Multimodal recognition of personality traits in social interactions. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 53–60 (2008)
13. Nihei, F., Nakano, Y.I., Hayashi, Y., Hung, H.H., Okada, S.: Predicting influential statements in group discussions using speech and head motion information. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 136–143 (2014)
14. Song, Y., Morency, L.P., Davis, R.: Multimodal human behavior analysis: learning correlation and interaction across modalities. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 27–30 (2012)
15. Hotelling, H.: Relations between two sets of variates. Biometrika **28**(3/4), 321–377 (1936)
16. Vinciarelli, A., Dielmann, A., Favre, S., Salamin, H.: Canal9: a database of political debates for analysis of social interactions. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–4 (2009)
17. Chatterjee, M., Park, S., Morency, L.P., Scherer, S.: Combining two perspectives on classifying multimodal data for recognizing speaker traits. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 7–14 (2015)
18. Qian, X., Wang, H., Zhao, Y., Hou, X., Hong, R., Wang, M., Tang, Y.Y.: Image location inference by multisaliency enhancement. IEEE Trans. Multimed. **19**(4), 813–821 (2017)
19. Zhang, S., Tian, Q., Hua, G., Huang, Q., Gao, W.: Generating descriptive visual words and visual phrases for large-scale image applications. IEEE Trans. Image Process. **20**(9), 2664–2677 (2011)
20. Kumar, V., Namboodiri, A.M., Jawahar, C.V.: Visual phrases for exemplar face detection. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 1994–2002 (2015)
21. Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling features for large scale partial-duplicate web image search. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 25–32 (2009)
22. Yang, X., Qian, X., Xue, Y.: Scalable mobile image retrieval by exploring contextual saliency. IEEE Trans. Image Process. **24**(6), 1709–1721 (2015)
23. Zhang, S., Yang, M., Wang, X., Lin, Y., Tian, Q.: Semantic-aware co-indexing for image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **37**(12), 2573–2587 (2015)
24. Okada, S., Hang, M., Nitta, K.: Predicting performance of collaborative storytelling using multimodal analysis. IEICE Trans. **99**(D(6)), 1462–1473 (2016)
25. Nakano, Y.I., Nihonyanagi, S., Takase, Y., Hayashi, Y., Okada, S.: Predicting participation styles using co-occurrence patterns of nonverbal behaviors in collaborative learning. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 91–98 (2015)

26. Vahdatpour, A., Amini, N., Sarrafzadeh, M.: Toward unsupervised activity discovery using multi-dimensional motif detection in time series, pp. 1261–1266 (2009)
27. Sanchez-Cortes, D., Aran, O., Jayagopi, D.B., Mast, M.S., Gatica-Perez, D.: Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. J. Multimodal User Interfaces **7**(1–2), 39–53 (2013)
28. Davis, J.W., Bobick, A.F.: The representation and recognition of human movement using temporal templates. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 928–934 (1997)
29. Turner, L.A., Perry, L.H., Sterk, H.M.: Constructing and Reconstructing Gender: The Links Among Communication, Language, and Gender. SUNY Press, Albany (1992)