Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2017) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

HCist 2017 - International Conference on Health and Social Care Information Systems and Technologies, 8-10 November 2017, Barcelona, Spain

# Design of a Knowledge-Based Agent as a Social Companion

Leo Wanner[a,b]1, Elisabeth André[c], Josep Blat[b], Stamatia Dasiopoulou[b], Mireia Farrús[b], Thiago Fraga[d], Eleni Kamateri[e], Florian Lingenfelser[c], Gerard Llorach[b], Oriol Martínez[b], Georgios Meditskos[e], Simon Mille[b], Wolfgang Minker[f], Louisa Pragst[f], Dominik Schiller[c], Andries Stam[g], Ludo Stellingwerff[g], Federico Sukno[b], Bianca Vieru, and Stefanos Vrochidis[e]

*[a]Catalan Institute for Research and Advanced Studies (ICREA); [b]Pompeu Fabra University, C/Roc Boronat, 138, 08018 Barcelona, Spain; [c]University of Augsburg, Universitätsstr. 6, 86159 Augsburg, Germany; [d]Vocapia Research 28, rue Jean Rostand Parc Orsay Université, 91400 Orsay, France; [e]Center for Research and Technologies Hellas (CERTH), 6th km Charilaou-Thermi Rd, P.O. Box 60361, 57001 Thermi, Thessaloniki, Greece; [f]University of Ulm, Albert Einstein Allee, 43, 89081 Ulm, Germany; [g]Almende BV, AK, Stationsplein 45, 3013 AK Rotterdam, The Netherlands*

**Abstract**

We present work in progress on an intelligent embodied conversation agent that is supposed to act as a social companion with linguistic and emotional competence in the context of basic and health care. The core of the agent is an ontology-based knowledge model that supports flexible reasoning-driven conversation planning strategies. A dedicated search engine ensures the provision of background information from the web, necessary for conducting a conversation on a specific topic. Multimodal communication analysis and generation modules analyze respectively generate facial expressions, gestures and multilingual speech. The assessment of the prototypical implementation of the agent shows that users accept it as a natural and trustworthy conversation counterpart. For the final release, all involved technologies will be further improved and matured.


Peer-review under responsibility of the scientific committee of the HCist - International Conference on Health and Social Care Information Systems and Technologies.

*Keywords:* conversation agent, dialogue, ontological knowledge model, multimodality, multilinguality, basic care, health care

* Corresponding author. Tel.: +34 93 542 2241; fax: +34 93 542 2517.
*E-mail address:* leo.wanner@upf.edu


Peer-review under responsibility of the scientific committee of the HCist - International Conference on Health and Social Care Information Systems and Technologies.

## 1. Introduction

Social companionship in the geriatric context is crucial[4,20,25,26]. As long as elderly were embedded into stable family structures which foresaw that several generations live together, this companionship was ensured at least to a certain degree. However, in the modern society, there is an increasing tendency among older people to live alone or in residences. As a result, the family bonds are at risk to weaken, such that elderly no longer feel sufficiently attended, lack affection and attention and are thus increasingly socially isolated. The problem is particularly significant for elderly migrants because they are often not integrated into the society of the host country either. Intelligent conversation agents could help in that they could converse, entertain, coach, etc. However, for this task, they must be versatile, eloquent, knowledgeable, and possess a certain cultural, social and emotional competence. Not all of these characteristics are displayed by state-of-the-art conversation agents. Many of them focus on the emotional (and, to a certain extent, social) competence[5,18,19,30,]. Hardly any are versatile, eloquent and knowledgeable. Versatility presupposes flexibility in dialogue conduction, but most of the agents follow a prescripted dialogue strategy and do not take into account that the course and content of a dialogue is also influenced by the culture of the human conversation counterpart. Eloquence presupposes full-fledged (multilingual) text generation, while most of the agents use predefined sentence templates for linguistic realization[1,29]. Knowledgeability presupposes a theoretically sound knowledge model over which the agent can reason and the possibility to acquire further knowledge if required by a conversation, while only very few agents are based on an ontological, expandable knowledge model and integrate an information search engine.

We attempt to account for some of the most significant limitations of the current state of the art in our work in progress on a flexible knowledge-based conversation agent. The agent is designed as an embodied companion for elderly migrants with language and cultural barriers in the host country and as a trusted information provider and mediator in questions related to basic care and healthcare.

## 2. Design of the Knowledge-Based Conversation Agent

The targeted agent (A) is expected to be able to conduct dialogues with users (U) as the following one:

**A**: *Why do you look so sad? What is wrong?*
**U:** *I feel sad; nobody came to see me already for two weeks.*
**A**: *Cheer up! What about going for a walk in the park?*
**U**: *How is the weather today?*
**A**: *The weather will be in general nice, but some showers cannot be excluded.*
*So don't forget to take an umbrella with you*
…

To have this capacity, the agent is envisaged to (i) be embedded into linguistic, cultural, social and emotional contexts of the users; (ii) be able to search for information in the web to either enrich its own knowledge repertoire or to offer to the user requested content; (iii) understand and interpret the multimodal (facial, gestural and multilingual verbal) communication signals of the user; (iv) plan the dialogue using ontology-based reasoning techniques according to the prior interpretation of the user signals; (v) communicate with the user with multimodal communication signals. The architecture of the agent in **Fig. 1** reflects these objectives.

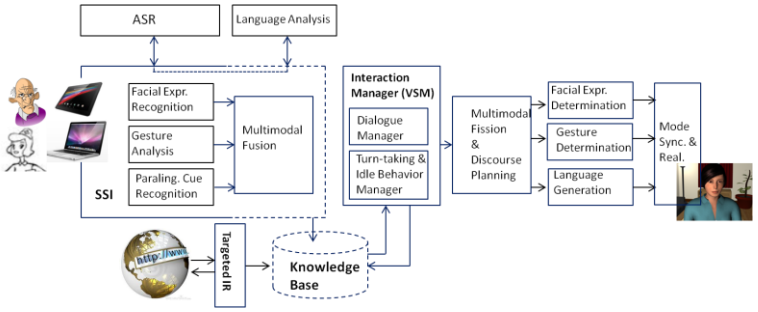


**Fig 1**: Architecture of the knowledge-based conversation agent

In this architecture, the communication analysis modules are controlled by the *Social Signal Interpretation* (SSI) framework[27]. The output produced by the communication analysis modules is projected onto genuine ontological (OWL) graphs, fused and stored in the knowledge base (KB), which also receives input from a search engine that retrieves background information from curated web sources. The dialogue manager (DM) is embedded into the *Visual Scene Maker* (VSM) framework[10], which, in addition, controls the agent's turn-taking and its non-verbal idle behavior[17]. The turn-taking strategy is based on a policy that determines whether the agent is allowed to interrupt the user's utterance and how it reacts when the user interrupts it. The agent displays idle behavior when it listens to the user, for example, mimicking the user's (positive) affective state or displaying different eye gazes[17].

The DM requests from the KB possible reasoned reactions to the communication move of the user and chooses the best in accordance with the analyzed move, the user's emotion and culture and the recent dialogue history. The OWL structures of the chosen reaction are passed by the DM to the fission and discourse planning module, which assigns to the content elements of these structures the realization modalities (voice, face, and/or body gesture) and plans their coherent and coordinated presentation. The three modality generation modules instantiate the form of the content elements assigned to them and their output is synchronized to ensure a coherent multimodal communication and realized.

Let us now briefly introduce the core modules. Verbal communication analysis implies speech recognition and language analysis. For speech recognition, VoxSigma (http://www.vocapia.com/) is used. Language analysis is realized as a sequence of processing stages that involves deep dependency parsing[2], rule-based graph transduction[6], and ontology design patterns[9]. Cf. **Fig. 2** for the intermediate structures of the ASR transcript *I feel sad*. Its knowledge graph representation is a declarative statement that contains an instantiation of the 'dul:Situation' class, which interprets the instances of ':CareRecipient' and ':Sad' classes as the experiencer and experienced emotion respectively of the event class ':Feel' instance (cf. **Fig. 3**).

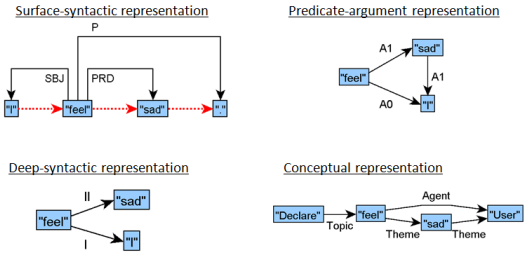


**Fig. 2**: Intermediate structures of language analysis

```
:declare a da:Declare ;
da:containsSemantics :feelCtx1 .
:feelCtx1 a dul:Situation ;
dul:includes :user1 ; dul:includes :sad1 ; dul:includesEvent :feel1 ;
dul:satisfies :feelDesc1 [a dul:Description] .
:feel1 a :Feel [rdfs:SubClassOf dul:Event] ;
dul:classifiedBy :Context [rdfs:SubClassOf dul:Concept] .
:sad1 a :Sad [rdfs:subClassOf :Emotion] ;
dul:classifiedBy :Theme [a dul:Concept] .
:user1 a :CareRecipient [rdfs:SubClassOf dul:Person] ;
dul:classifiedBy :Experiencer [rdfs:SubClassOf dul:Concept] .
```

**Fig. 3**: OWL structure of the statement 'I feel sad'

Facial expression analysis, gesture recognition and paralinguistic cue analysis focus so far on the identification of the affective state of the user. For facial expression analysis, *Action Units* (AUs) from the *Facial Action Coding System*[8,23] are used. To determine the AUs, first SIFT-based features are extracted from automatically detected facial landmarks, and subsequently linear classifiers are applied to assign probabilities to the targeted AUs. During gesture analysis, the agitation of the user in terms of hand movements is detected using video frame filter masks. To obtain paralinguistic affective cues, we use Wagner et al.'s model[27]. Extracted facial, gestural and paralinguistic cues are combined using event fusion strategies[15] and projected onto values in the valence/arousal space, which are then represented in the KB.

The agent's KB contains ontologies that cover: (i) models for the representation, integration and interpretation of the content of the user's multimodal communication; (ii) background knowledge and user profile and behavior pattern ontologies; and (iii) healthcare and medical ontologies[22]. The knowledge integration and interpretation models define how the structures can be combined to derive high-level interpretations. For this, a lightweight ontology schema models the types of relevant structures and their interpretation strategies by the reasoner. **Fig. 4** depicts the vocabulary for the interpretation of the user's statement *I feel sad* and the complementary information from the visual input 'low mood' encoded in terms of valence/arousal values[21]. As can be observed, the ontology extends the 'leo:Event' concept of LODE[24], taking advantage of existing vocabularies for description of events and observations. The relation between observation types and context is modeled in terms of the 'Context' class. This

allows for the introduction of one or more 'contains' property assertions that refer to observations. The fact that the user is sad constitutes contextual information that is modeled as an instance of 'Context', which is further associated with an instance of 'Sad'.

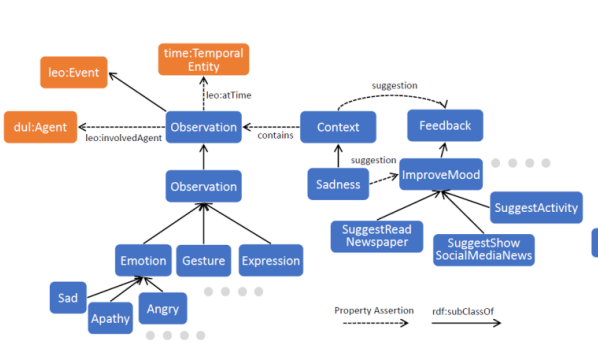


**Fig. 4**: Observation and Context models

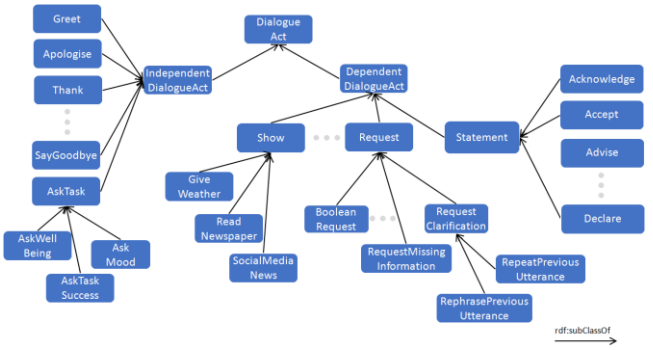


**Fig.5**: Excerpt of the dialogue acts ontology

**Fig. 4** furthermore contains an excerpt of the domain ontology used to infer feedback and suggestions based on the emotional state of the user. For each context, one or more 'suggestion' property assertions can be defined and associated with feedback instances that can improve user's mood. In our example, 'Sadness' is a subclass of 'Context', defined in terms of the following equivalence axiom: $Sadness \equiv Context \cap \exists contains.Sad$. It also defines the property restriction $Sadness \subseteq \exists suggestion.ImproveMood$, which specifies the type of feedback needed when this emotional context is detected. The 'ctx1' instance is classified in the 'Sadness' context class, which further inherits the restriction about the potential feedback that could be given to improve the mood of the user. All three subclasses of the 'ImproveMood' concept are retrieved and sent back to the DM as possible system actions. The DM chooses the one that is to be communicated to the user. The choice is grounded in the dialogue act typology shown in **Fig. 5** as well as in the user's emotion and culture and the recent dialogue history. To avoid the predefinition of all user and system actions and be able to handle arbitrary input from both the language analysis and the KB, the choice is defined for general features such as the respective dialogue act and the topics, constituted by the classes associated with the possible system actions. For instance, the three possible system actions from above share the dialogue act 'Statement'. However, the topics differ. Thus, the first action has the topics 'newspaper' and 'read', the second 'socialmedia' and 'read', and the third 'activity'. Individuals from a collectivistic culture tend to be more tightly integrated in their respective social groups, while individuals from an individualistic culture less so[12]. Therefore, the DM would propose to the user with a collectivistic culture background to read aloud news from social media, and select one of the other options if the user's culture is individualistic.

Once the appropriate system action has been determined by the DM, the fission module assigns to the individual mode generation modules the content elements from the OWL graph that are to be expressed by the respective mode. Language generation follows the inverse cascade of processing stages depicted for analysis; see **Fig. 2** above. The language generator consists of multilingual rule-based[28] and statistical[3] graph transduction components. As speech generator, which takes as input the output provided by the language generator, we use CereProc's TTS (https://www.cereproc.com/). In parallel to the cascaded proposition realization model, a hierarchical prosodic model is deployed, which captures prosody as a complex interaction of acoustic features at different phonological levels in the utterance (i.e., prosodic phrases, prosodic words and syllables) and the information structure[7].

For its non-verbal appearance, the agent is realized as an embodied virtual character. Cultural gestures and facial expressions are generated according to the semantics of the message that is to be communicated. To facilitate the required variety of facial expressions and avoid manual design of all possible expressions for each character, the valence-arousal representation of emotions in the continuous 2D and 3D space is used[11,13,14]. Because of its parametric nature, the valence-arousal space can be easily applied to a variety of faces.

## 3. Evaluation

The assessment of the quality of the first prototypical implementation of the agent has been carried out in the context of three different use cases. In the first, it acts as a social companion of elderly with German respectively Turkish background, in the second it acts as an assistant of Polish care givers who take care of German elderly, and in the third it is supposed to be a healthcare adviser of migrants with North African background. Qualitative evaluation trials have been run with respect to the agent's appearance, trustworthiness, competence, naturalness, friendliness, speech and language understanding and production quality, etc. As far as the agent's appearance is concerned, it was rated as still to be too rigid and unnatural, which was to be expected given the preliminary design of the virtual character that embodies the agent. When asked whether the agent was proactive in addressing the user and whether the communicative goal of the agent was in general clear, better marks were achieved (3.23 respectively 3.25 on a five value Likert scale, with '1' being the worst and '5' being the best). Also, most of the evaluators agreed that the agent provides the right amount of information when being asked (3.27 on the Likert scale). In general, it can be thus assumed that even in its preliminary appearance the agent is considered to be a competent conversation partner.

For further illustration of the performance of the 1st prototype of the agent, Table 1 displays an excerpt of the questionnaire on the quality of language production by the agent.

**Table 1**: Evaluation of the language production competence of the 1st prototype of the agent: '1'= "disagree"; '5' = "compeletely agree"

| **Evaluation statement** | **Likert scale value (SD)** |
|---|---|
| The voice of the agent sounds natural | 2.81( ± 1.33) |
| The voice of the agent is expressive | 2.24 ± 1.12 |
| The statements uttered by the agent are perfectly understandable | 3.19 (± 0.98) |
| The language as used by the agent was perfectly grammatical | 2.77 (±1.12) |
| The agent expresses itself accurately | 3.45 (±1.03) |
| The agent talks coherently (in the case of a multi-sentential statement) | 3.43 (± 0.97) |

As can be observed, the voice of the agent still needs to be improved. In particular, the prosody of the agent is perceived to be monotonous in the case of a multi-sentential discourse or when reading a newspaper. We are currently about to experiment with novel techniques for prosody enrichment. The grammaticality has also been rated as deficient, while the content of the discourse is already considered to be rather accurate. Obviously, all technologies still need to be improved considerably.

## 4. Conclusions

We presented the design and a preliminary evaluation of the first prototype of an intelligent embodied conversation agent, which is aimed to conduct socially competent and culture-sensitive multilingual conversations in the context of basic care and healthcare. In its current version, the agent is able to understand and communicate in German, Polish, and Spanish; in its next version, the language competence will be extended to cover also Arabic and Turkish. The results of the first round of evaluation is encouraging, although some clear need for further improvement of the individual technologies has been identified. Currently, the identified shortcomings are about to be addressed.

## Acknowledgments

The presented work is funded by the European Commission as part of the H2020 Program under the contract number 645012. We would like to thank our colleagues Chiara Baudracco, Jutta Mohr, Eylem Ög, Valérie Sarholz, and Benjamin Schäfer for organizing and carrying out the evaluation trials and for their patience with the numerous inadequacies of KRISTINA.

## References


1. Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssadou, E., Damian, I., Ennis, C., Egges, A., Gebhard, P., Jones, H., Ochs, M., Pelachaud, C., Porayska-Pomsta, K., Rizzo, P., Sabouret, N.: The TARDIS framework: Intelligent virtual agents for social coaching in job interviews. In: Reidsma, D., Katayose, H., Nijholt, A. (eds.) ACE, vol. LNCS, 8253. 2013; p. 476–491. Springer, Heidelberg.
2. Ballesteros, M., Bohnet, B., Mille, S., Wanner, L.: Data-driven deep-syntactic dependency parsing. *Natural Language Engineering*. 2016; **22(6)**:939–974.
3. Ballesteros, M., Bohnet, B., Mille, S., Wanner, L.: Data-driven sentence generation with non-isomorphic trees. In: *Proceedings of the Conference of the NAACL: Human Language Technologies*; 2015. p. 387–397.
4. Baldassare, M., Rosenfield, S., and Rook, K. The types of social relations predicting elderly well-being. *Res on Aging*. 1984. **6(4)**:549 – 559.
5. Baur, T., Mehlmann, G., Damian, I., Gebhard, P., Lingenfelser, F., Wagner, J., Lugrin, B., André E.: Context-Aware Automated Analysis and Annotation of Social Human-Agent Interactions. ACM Transactions on Interactive Intelligent Systems. 2015; **5(2)**.
6. Bohnet, B., Wanner, L. Open source graph transducer interpreter and grammar development environment. In: *Proceedings of the International Conference on Language Resources and Evaluation*; 2010.
7. Domínguez, M., Farrús, M., Burga, A., Wanner, L.: Using hierarchical information structure for prosody prediction in content-to-speech application. In: *Proceedings of the 8thInternational Conference on Speech Prosody*; 2016.
8. Ekman, P., Rosenberg, E.L. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA; 1997.
9. Gangemi, A.: The Semantic Web. In: *Proceedings of the 4th International Semantic Web Conference*; 2005, p. 262 – 27
10. Gebhard, P., Mehlmann, G.U., Kipp, M.: Visual SceneMaker: A Tool for Authoring Interactive Virtual Characters. *Journal of Multimodal User Interfaces: Interacting with Embodied Conversational Agents*, Springer-Verlag. 2012; **6(1-2)**:3– 11.
11. Gunes, H., Schuller, B.: Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 2013; **31(2)**:120–136.
12. Hofstede, G.H., Hofstede, G. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage. 2001.
13. Hyde, J., Carter, E.J., Kiesler, S., Hodgins, J.K.: Assessing naturalness and emotional intensity: a perceptual study of animated facial motion. In: *Proceedings of the ACM Symposium on Applied Perception*. 2014; p 15–22. ACM.
14. Hyde, J., Carter, E.J., Kiesler, S., Hodgins, J.K.: Using an interactive avatar's facial expressiveness to increase persuasiveness and socialness. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015; p. 1719 – 1728. ACM.
15. Lingenfelser, F.,Wagner, J., André, E., McKeown, G., Curran, W.: An event driven fusion approach for enjoyment recognition in real-time. In: *Proceedings of the Multimedia Conference*. 2014; p. 377–386.
16. Mehlmann, G., Janowski, K., André, E.: Modeling Grounding for Interactive Social Companions. *Journal of Artificial Intelligence: Social Companion Technologies*. 2016. **30(1)**:45–52.
17. Mehlmann, G., Janowski, K., Baur, T., Häring, M., André, E., Gebhard, P. Exploring a Model of Gaze for Grounding in HRI. In: *Proceedings of the 16th International Conference on Multimodal Interaction*. 2014; p. 247–254. ACM.
18. Ochs, M., Pelachaud, C.: Socially Aware Virtual Characters: The Social Signal of Smiles. *IEEE Signal Processing Magazine*. 2013; **30(2)**:128–132.
19. Pfeifer Vardoulakis, L., Ring, L., Barry, B., Sidner, C., Bickmore, T.: Designing relational agents as long term social companions for older adults. In: *Proceedings of the 12th International Conference on Intelligent Virtual Agents*. 2012.
20. Pickett Y, Raue, PJ, Bruce, ML. Late-life depression in home healthcare. *J Aging Health* 2012; **8(3)**: 273–284.
21. Posner, J., Russell, J., Peterson, B.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development and psychopathology. *Development and psychopathology*. 2005; **17(3)**.
22. Riaño, D., Real, F., Campana, F., Ercolani, S., Annicchiarico, R.: An ontology for the care of the elder at home. In: *Proceedings of the 12th Conference on Artifcial Intelligence in Medicine: Artifcial Intelligence in Medicine*. 2009; p. 235–239. AIME '09, Springer-Verlag, Berlin.
23. Savran, A., Sankur, B., Bilge, M.T.: Regression- based intensity estimation of facial action units. *Image and Vision Computing* 2012; **30(10)**:774 –784.
24. Shaw, R., Troncy, R., Hardman, L.: Lode: Linking open descriptions of events. In: *Proceedings of the 4th Asian Conference on the Semantic Web*. 2009; p. 153–167. Shanghai, China.
25. Sorkin, D., Rook, K.S. and Lu, J.L.: Loneliness, lack of emotional support, lack of companionship, and the likelihood of having a heart condition in an elderly sample. *Ann. Behav. Med.* 2002. **24**: 290–298.
26. Vlachantoni, A., Shaw, R., Willis, R., Evandrou, M., Falkingham, J., Luf, R.. Measuring unmet need for social care amongst older people. *Population Trends*. 2011; **145**:1–17.
27. Wagner, J., Lingenfelser, F., André, E.: Building a robust system for multimodal emotion recognition. *Emotion Recognition: A Pattern Analysis Approach*. 2015; p. 379–419. John Wiley & Sons, Hoboken, NJ.
28. Wanner, L., Bohnet, B., Bouayad-Agha, N., Lareau, F., Nicklass, D.: MARQUIS: Generation of user-tailored multilingual air quality bulletins. *Applied Artifcial Intelligence*. 2010; **24(10)**:914–952.
29. Yasavur, U., Lisetti, C., Rishe, N.: Lets talk! Speaking virtual counselor offers you a brief intervention. *Journal of Multimodal User Interfaces*. 2014; **8(4)**:381–398.
30. Zeng, Z., Pantic, M., Roisman, G., Huang, T.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*. 2009; **31(1)**:39–58.