# BRIDGING COMPUTER VISION AND SOCIAL SCIENCE : A MULTI-CAMERA VISION SYSTEM FOR SOCIAL INTERACTION TRAINING ANALYSIS

*Jixu Chen* *Ming-Ching Chang* *Tai-Peng Tian* *Ting Yu* *Peter Tu*

GE Global Research, Niskayuna NY USA

## ABSTRACT

We investigate the use of a vision-based system capable of estimating social states such as rapport and hostility. We study the correlation between interpretations automatically generated by our system and those reported by social scientists. Our multi-camera vision system collects visual cues including location (proximity), motion, pose, gaze, and facial expressions in real-time from multiple subjects moving freely in an unconstrained environment. We performed experiments on 80+ subjects. Preliminary regression analysis suggests high correlation between machine distilled time series signals and assessments made by human experts.

***Index Terms***— social interaction, social signal processing, gaze, facial expression, body pose, behavior analysis.

## 1. INTRODUCTION

It is well known in the social sciences that *non-verbal cues* such as facial expressions and body gestures embody important information [1, 2]. Social signal processing (SSP) [3] is a research field that focuses on enabling computers to interpret non-verbal cues and recognize human social interactions automatically. Computer vision techniques such as facial expression detection [4], gaze estimation [5, 6] and gesture recognition [7] have been successfully applied to social interaction analysis. However, there exists two major challenges in the development of SSP : (1) Non-verbal cues are hard to collect using automated methods, as traditionally they are only available from manual annotation, which is costly and labor intensive [8, 9]. (2) Studies in computer vision and social science often do not share a common terminology. Thus the interpretation of the term "non-verbal cues" may not necessary mean the same thing to the two communities. In this paper, we focus on a set of scenario-based experiments with an emphasis on social skill training.

We investigate the use of a computer vision system (Fig.1) for evaluating social communications and interactions that take place during task-oriented scenarios. Social awareness and the ability to understand and manage social signals are at the core of social intelligence. Proficiency in such social skills is often the key for success in practical scenarios such as negotiation, interviewing, marketing, etc. Our aim is to improve training by considering social signals ranging from body orientation, movement, gaze, expression, and pose as depicted in Fig.2(a). We propose to facilitate the training course by leveraging an automatic visual system to: (1) provide on-line social interaction evaluation and (2) investigate the correlation between machine distilled visual cues and assessments obtained from social scientists.

We assume that by jointly considering visual behavioral cues from trainees and role-players, it is possible to predict (in a completely automated fashion) assessments made by social scientists. The main contributions of this paper include: (1) real-time end-to-end system capable of predicting various social interaction states (i.e., rapport) based on visual cues (Section 2); and (2) experimental validation of this system using real data (Section 3). Our findings will help bridge the gap between machine vision and the social sciences.

## 2. A MULTI-CAMERA SYSTEM FOR SOCIAL INTERACTION ANALYSIS

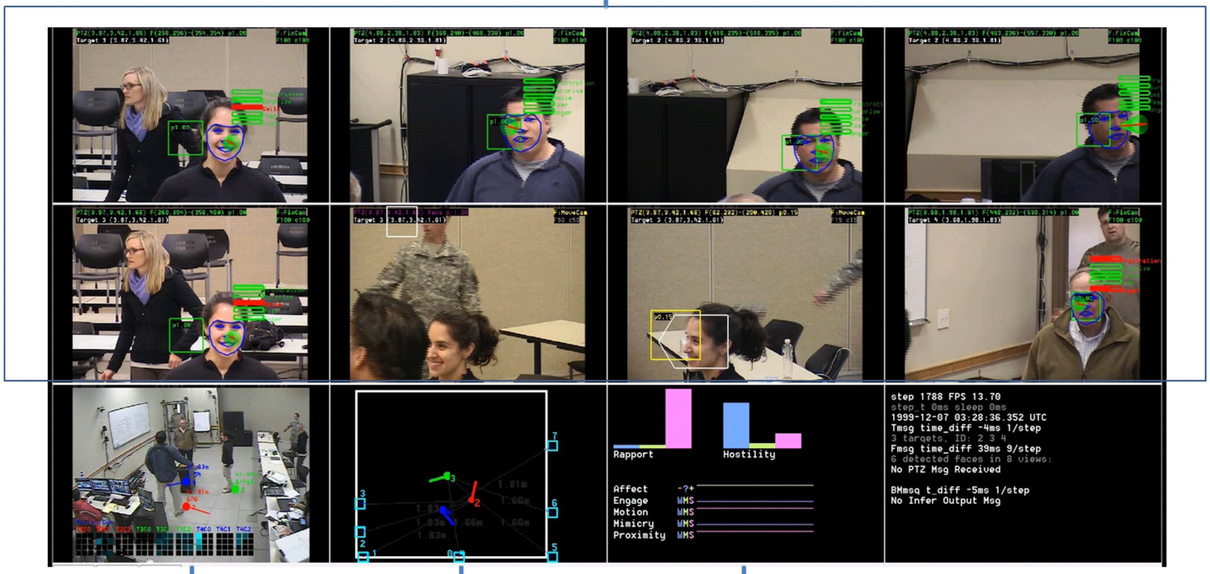
As part of a social interaction training course, trainees participated in a set of challenging role-play scenarios, which are designed to assess their communication and decision making skills. During each exercise, both the trainee and role-players can operate freely. Due to occlusion and rapid movements, it is difficult to capture certain visual cues (especially facial shots) in a reliable fashion even with the use of multiple fixed cameras. To this end, we have developed a multi-view active Pan-Tilt-Zoom (PTZ) camera system that addresses these issues.

### 2.1. System Operation

Our computer vision system consists of a set of fixed RGB+D cameras and a ring of PTZ cameras. As individuals enter the site they are tracked using the fixed cameras. PTZ cameras are then automatically targeted onto individual faces in the scene. Fixed cameras provide position tracking, body pose and mo-

This work was supported by The Defense Advanced Research Projects Agency (Contract D13PC00002). The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

8 PTZ Cameras are automatically targeted on to tracked individuals resulting in expression and gaze estimates. Expressions include: joy, frustration, anger, surprise and fear



```
step 1788 FPS 13.70
step_t 0ms sleep 0ms
1999-12-07 03:28:36.352 UTC
Tmsg time_diff -4ms 1/step
3 targets, ID: 2 3 4
Fmsg time_diff 39ms 9/step
6 detected faces in 8 views:
No PTZ Msg Received

BMmsq t_diff -5ms 1/step
No Infer Output Msg
```

Multi-view tracking and articulated motion analysis

Automatic PTZ assignment and control

Time series representation of various social signals as well as an estimation of group level social states such as rapport and hostility.

**Fig. 1**. **Overview of the proposed system running in a social interaction training course.**

tion information. PTZ cameras characterize each individual's gaze direction and facial expressions including *anger*, *fear*, *joy*, *surprise*, and *frustration* on a per-frame basis in real-time. Details of our system are described in [10] and [11]. Fig. 1 illustrates the system in action during a training session.

After each training session, both the video recording and machine generated visual cues are saved to a central database. Social scientists then independently rate the observed level of rapport (from 1 to 5) by observing raw videos. Judgments are made according to the non-verbal cues of Bernieri et.al. [9], which include: (1) *Expressivity* – the extent to which interactants' total behavior is active, animated, and exaggerated; (2) *Synchrony* – the extent to which the behavior of each interactant was similar to, and coordinated with, each other; (3) *Proximity* – average distance separating the interactants' noses and closest knees.

Note that non-verbal cues commonly annotated by social scientists are different from the visual cues generated from a machine vision system, although they are highly correlated in general. A significant contribution of this work is that we derive a data-driven regression model to characterize such relationships (details given in Section 2.2). Fig. 2(b) depicts the relationship between the machine-generated visual cues and social science annotations.

### 2.2. Regression from Visual Cues to Rapport Estimation

Compared to existing social signal processing systems [12–14] where no specific distinction of roles was emphasized (particularly in distinguishing the trainee from role-players), we introduce two important considerations in designing the rapport regressor. (1) Interaction between the trainee and role-players, which reflects the social skills of the trainee, is our primary aim. We focus on the trainee and lower the effect of interactions between role-players. (2) The temporal resolution associated with machine distilled visual cues and social sciences are different. Machine generated visual cues are on a per-frame basis, where the annotations from social scientists consider whole training session in their entirety. Therefore sequence-level visual cues must be derived from the integration of per-frame measures.

By considering a large set of aggregate statistics gleaned from each sequence, we derive a feature vector $V$ that can be used to characterize each training sequence. These statistics are based on the results of tracking, motion and facial analysis.

Given the training sequences annotated by social scientists, a simple linear regressor $F(V) = W \cdot V$ is employed to predict the level of rapport $R$. A linear coefficient matrix $W$ can be obtained at the learning stage, by minimizing the prediction error in the training data:

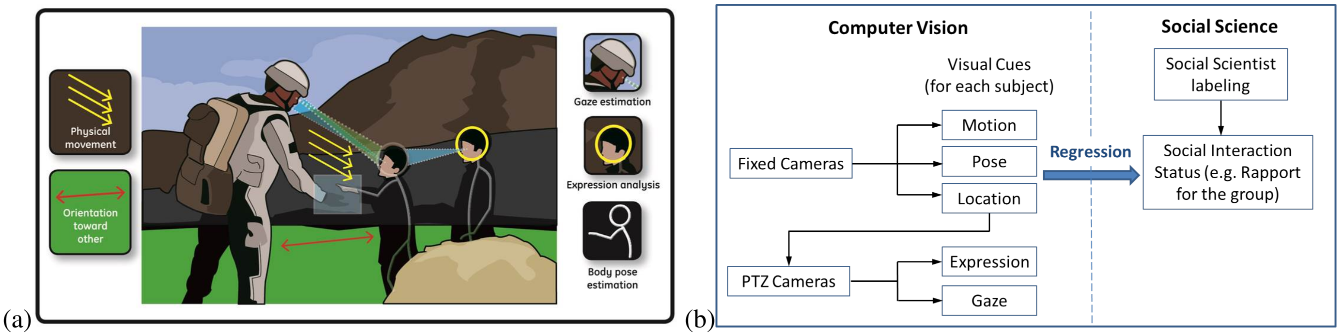


**Fig. 2**. (a) **A social interaction scenario in the wild.** (b) **Proposed system module diagram.**

$W = \arg\min_W \sum_{i=1}^{N} \|F(V_i) - R_i\|^2$, where $N$ is the number of sequences. The minimization is solved using a standard SVD approach.

## 3. EXPERIMENTAL RESULTS

We first evaluate the visual cue estimation performance of our machine vision system in terms of tracking and expression analysis. We then evaluate the performance of the rapport estimation method.

**Initial evaluation of facial expression recognition** is performed on enacted facial expressions of ten subjects at our test site. Each subject is asked to perform various expressions: anger (**An**), fear (**Fe**), joy (**Jo**), surprise (**Su**) and frustration (**Fr**). The models were developed based on CK+ facial expression database [15, 16]. For ‘frustration’ which is not include in CK+, we consider a two-fold cross evaluation using our in-house data-set. Table 1 shows the facial expression classification performance. We set the threshold of each expression detector to be slightly higher than the learned threshold. Therefore minor facial expressions might be classified as neutral (**Ne**).

**Evaluation of person tracking and face analysis** on the live system is performed by manual validation. 36 training sessions of 40 trainees and role-players are used in this evaluation. Each training session lasts between 2.5 and 8 minutes. Reliable person tracking is crucial to the evaluation, since errors caused by lost track or ID switch can degrade performance significantly. We found that the tracker only failed in two training sessions (error rate $= 5.6\%$). Since it is very time consuming to manually label all 1280 minutes of videos from eight PTZ cameras, we selected 1072 frames for manual labeling with respect to facial features. One of three gaze directions (frontal, left, right) and one of five facial expressions (frustration, smile, surprise, fear, anger) are labeled. Using the manual labeling as ground-truth, we found that $2.2\%$ of the frames exhibited false gaze estimate and $1.9\%$ exhibited false expression estimate.

**Evaluation of Rapport Estimation.** The dataset for this evaluation was collected over 10 days with 80 trainees. Social interaction performance data for each trainee was evaluated by social scientists. We use the rapport scores from social science annotations as ground truth. Recordings from 40 trainees were used to learn a linear regressor (in Section 2.2), which was tested on the recordings from the remaining 40 trainees. The correlation between this prediction and ground-truth was $0.35$ ($p < 0.05$), which indicates a significant positive correlation. For each session, two sets of annotations are available from different social scientists. The correlation between these two annotations was $0.68$. Although the machine generated annotation is currently not as consistent as the expert’s, this preliminary result is promising for social interaction analysis. This benchmark result provides valuable insight for future studies.

| | Ne | An | Fe | Jo | Su | Fr |
|---|---|---|---|---|---|---|
| Ne | 99.0 | 0.3 | 0.0 | 0.2 | 0.0 | 0.5 |
| An | 16.3 | 76.3 | 0.0 | 0.0 | 0 | 7.4 |
| Fe | 22.4 | 0.7 | 55.9 | 19.6 | 0.7 | 0.8 |
| Jo | 16.5 | 1.4 | 1.0 | 81.1 | 0.0 | 0.0 |
| Su | 19.5 | 0.0 | 0.0 | 0.0 | 80.5 | 0.0 |
| Fr | 15.6 | 2.5 | 0.0 | 0 | 0.0 | 81.8 |

**Table 1**. **Confusion matrix** of expression recognition results.

## 4. CONCLUSION

Social interaction analysis has been studied for many decades. Recently, computer vision techniques have been applied to this area of investigation. In this work, we perform a joint analysis of the findings from both computer vision and the social sciences. We study and compare the findings between the two fields. Our preliminary experimental results show that machine vision systems can be used for automatic estimation of valuable social behavior states.

**Acknowledgment.** We gratefully acknowledge Zachary Horn and Kent Halverson from the Aptima group for providing valuable behavior labeling data for this study.

## 5. REFERENCES

[1] Virginia P Richmond, James C McCroskey, and Mark Hickson, *Nonverbal behavior in interpersonal relations*, Prentice Hall Englewood Cliffs, NJ, 1991.

[2] Mark Knapp, Judith Hall, and Terrence Horgan, *Nonverbal communication in human interaction*, Cengage Learning, 2013.

[3] Alessandro Vinciarelli, Maja Pantic, and Herve Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, pp. 1743–1759, 2009.

[4] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[5] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt, "Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2001, CHI '01, pp. 301–308.

[6] Dan Witzner Hansen and Qiang Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 3, pp. 478–500, 2010.

[7] Dimitris Metaxas and Shaoting Zhang, "A review of motion analysis methods for human nonverbal communication computing," *Image and Vision Computing*, vol. 31, no. 67, pp. 421 – 433, 2013.

[8] Jon E Grahe and Frank J Bernieri, "The importance of nonverbal cues in judging rapport," *Journal of Nonverbal behavior*, vol. 23, no. 4, pp. 253–269, 1999.

[9] Frank J Bernieri, John S Gillis, Janet M Davis, and Jon E Grahe, "Dyad rapport and the accuracy of its judgment across situations: A lens model analysis.," *Journal of Personality and Social Psychology*, vol. 71, no. 1, pp. 110, 1996.

[10] Ming-Ching Chang, Jixu Chen, Tai-Peng Tian, Peter Tu, and Ting Yu, "Social interaction analysis," in *NSF/FBI/DARPA Workshop on Frontiers in Video and Image Analysis*, Washington DC, USA, 2014.

[11] Ming-Ching Chang, Jixu Chen, Tai-Peng Tian, Ting Yu, and Peter Tu, "A live video social interaction analysis system," Manuscript submitted to ICIP 2015, 2015.

[12] Tian Lan, Yang Wang, Weilong Yang, S.N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 8, pp. 1549–1562, Aug 2012.

[13] D.B. Jayagopi, H. Hung, Chuohao Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 3, pp. 501–513, March 2009.

[14] Daniel Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[15] Takeo Kanade, Jeffrey Cohn, and Ying-Li Tian, "Comprehensive database for facial expression analysis," in *IEEE Auto. Face Gesture Recognition*, March 2000, pp. 46 – 53.

[16] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *CVPRW*. IEEE, 2010, pp. 94–101.