Ubiquitous and Mobile Computing
CS 528: *Insert Topic*

Your Reactions Suggest You Liked the Movie: Automatic Content Rating via Reaction Sensing

------Naihui Wang, Qiuzhe Ma

Computer Science Dept.

Worcester Polytechnic Institute (WPI)

Motivation

- **Conventional systems:**

1. Today's ratings are most often a simple number, that often leaves the new user asking for more.
2. Eliciting a carefully considered rating from users is difficult, partly due to the lack of incentives.

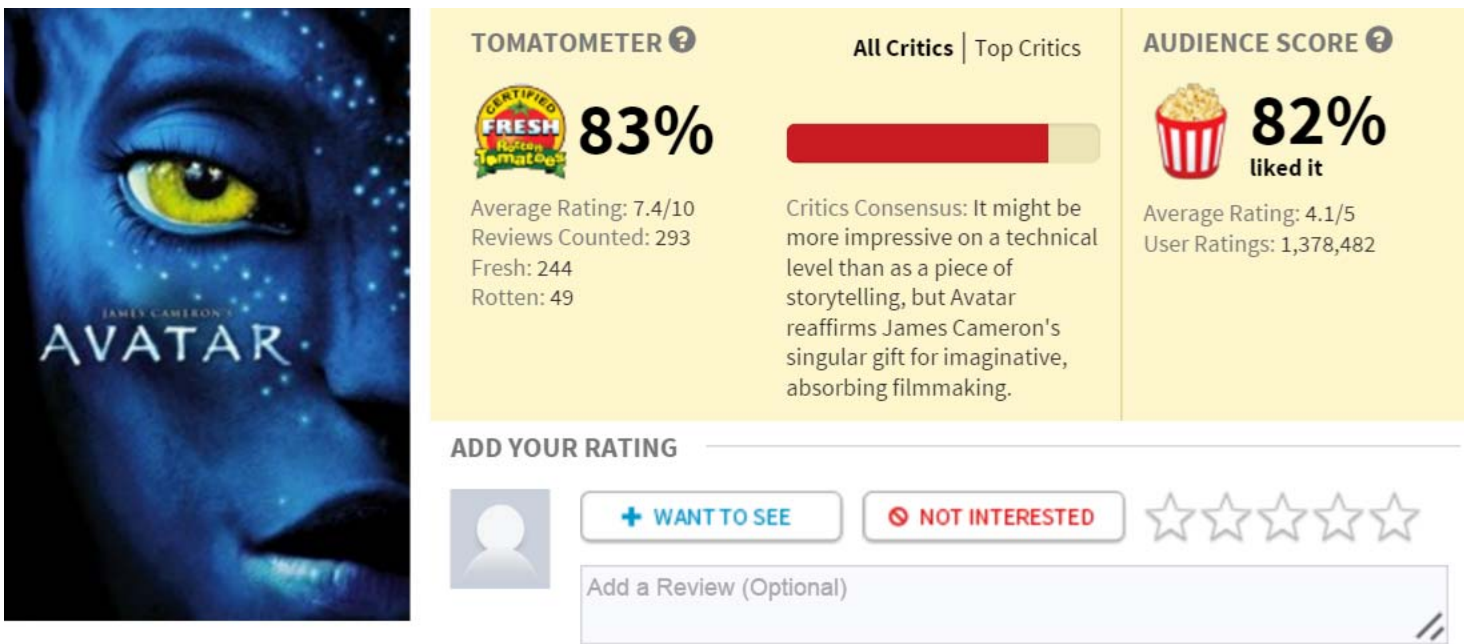


Figure 1: Rating of Avatar from rotten tomatoes

Motivation

- **Key observation** : The rich set of sensors available on today's smartphones and tablets could be used to capture a wide spectrum of user reactions while users are watching movies on these devices.

- **Goal :** Content rating systems of the future will require minimal user participation and yet provide rich, informative ratings.

- **Result of this paper:** This paper makes an attempt to realize a system called Pulse, which can learn the mapping between the sensed reactions and ratings, then automatically compute users' ratings.

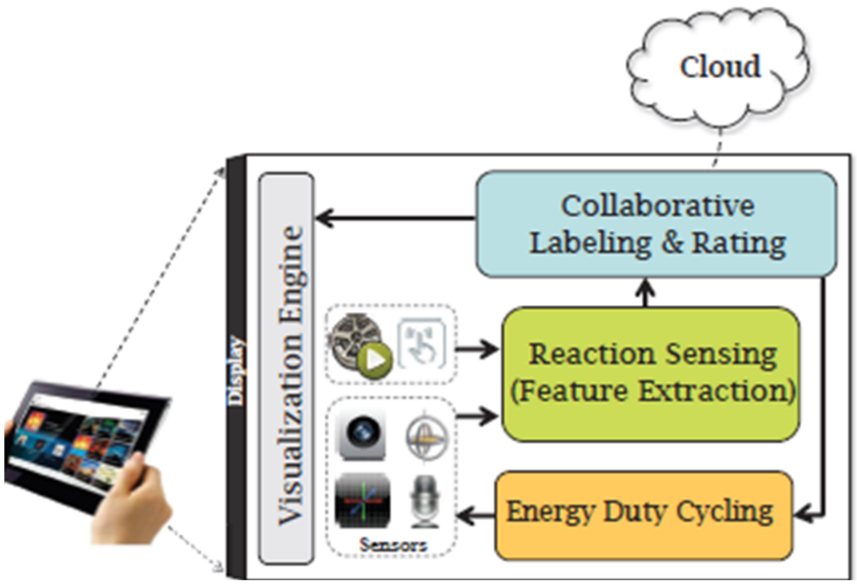
Functions of Pulse:

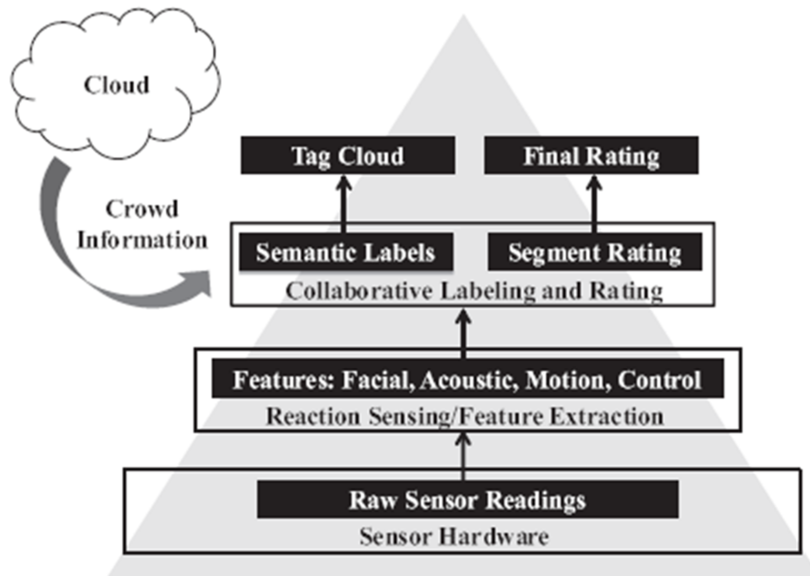
- The timeline of a movie can be annotated with reaction labels (e.g., funny, intense, warm)
- Senses user reactions and translates them to an overall system rating.

Figure 2. Envisioned movie ratings for the future – a conventional 5-star rating; a tag-cloud of user reactions; movie clips indexed by these reactions.

SYSTEM OVERVIEW

- **Main modules :** Reaction Sensing and Feature Extraction (RSFE), Collaborative Labeling and Rating (CLR), and Energy Duty-Cycling (EDC).





- **RSFE:** processes the raw sensor readings and extracts features to feed to CLR.
- **CLR:** The CLR module processes each (1 minute) segment of the movie to create a series of "semantic labels" as well as "segment ratings". Finally, the segment ratings are merged to yield the final "star rating " while the semantic labels are combined to create a tag-cloud.
- **EDC:** EDC ' s task is to minimize the energy consumption due to sensing.

System design: RSFE

- **Visual:** Pulse detects the face through the tablet camera, detects the eyes using blink detection, and finally tracks the key points.

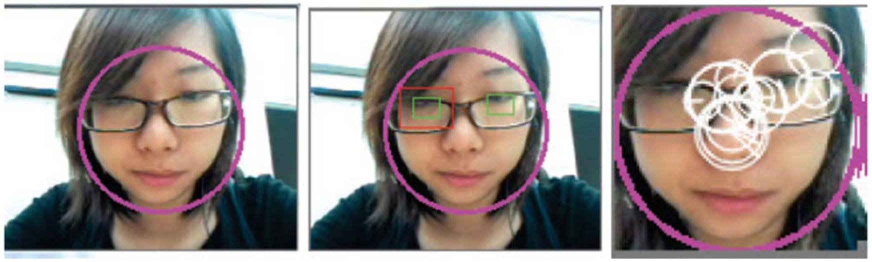


Figure 3: Visual sensing in Pulse: Face, eye, and blink detection for a user with spectacles.

- **Acoustic:**

1. **Voice Detection:** The recorded sounds drop sharply at around 4kHz. At less than 4kHz, the movie soundtrack with and without human voice are comparable, and therefore non-trivial to separate.
2. **Laughter Detection:** Pulse assumes that acoustic reactions during a movie are either speech or laughter – so, once human voice is detected, it needs to classified to one of the two categories. We use a support vector machine (SVM) and train it on the Mel-Frequency Cepstral Coefficients (MFCC) as the principle features.

- **Touch Screen:** Users tend to skip boring segments of a movie and, sometimes, may roll back to watch an interesting segment again.

System Design: CLR

- **Ratings:** Pulse employs Collaborative Filtering and Gaussian Process Regression (GPR) to cope with such ambiguities (detailed later). To convert segment ratings to the final rating, Pulse uses a weighted averaging function.

- **Labels:** Semantic labels are English labels assigned to each segment of the movie. CLR generates two types of such labels :

1. Reaction labels are direct outcomes of reaction sensing, reflecting on the viewer's raw behavior while watching the movie (e.g., laugh,smile focused , distracted, nervous, etc.).
2. Perception labels reflect on subtle emotions evoked by the corresponding scenes (e.g., funny, exciting, warm, etc.) Pulse employs a semi-supervised learning method combining Collaborative Filtering and SVM to predict perception labels.

Experiment Methodology

- 11 volunteers, 6 new movies, use Pulse video player, watch at any time and place
- After watch: rate segments, perception label, final "star" rating

Challenges

Predicting human judgment, minute by minute, is quite difficult.

- **Heterogeneity in users behavior**
  Eg: result detected: *Hold device still*
  reason: *Movies are intense VS. boring*

- **Heterogeneity in environment factors**
  Eg: *Watch in the office VS. at home*

- **Heterogeneity in user tastes**
  Eg: *Normal VS. hilarious*

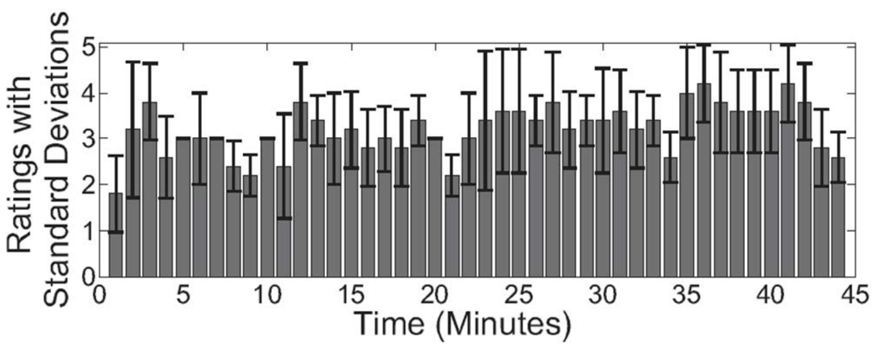


Figure 11. High Std. Dev. in ratings across users.

Resolving Challenges
Pulse's Learning Approach

Although users exhibit heterogeneity overall, their reactions to certain parts of the movie are coherent.
Analyze the **collective behavior** of multiple users to **extract** only these **coherent signals**.

- **For Segment Ratings:**
  Combine collaborative filtering with Gaussian Process Regression (GPR)

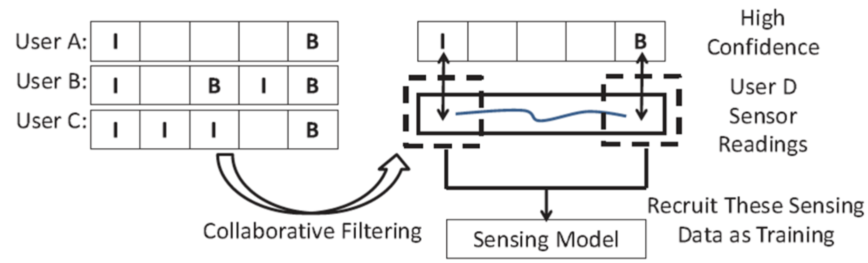


Figure 12. Pulse learns a custom model from high-confidence segments.

- **For perception labels**:
  Combine collaborative filtering with support vector machines (SVM)

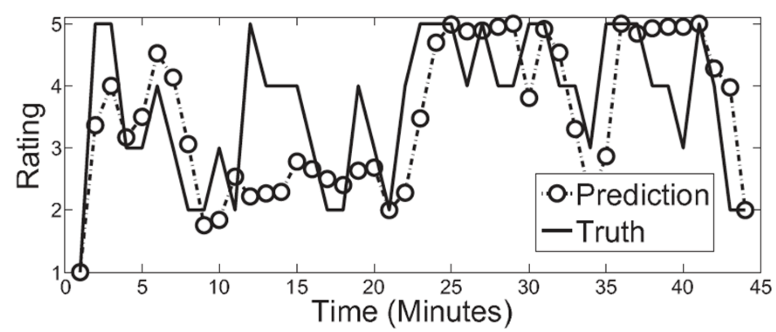


Figure 13. Collaborative filtering and GPR improve prediction – circles are the Pulse's predictions

Additional Challenges

- **Time-scale of Ratings:**

  mismatch between time-scale of sensed reactions and time-scale of human ratings.

  Eg: *a laughter lasts a few seconds VS. human rating one for each minute.*

  Resolution: 3 second window; Aggregate back to minute granularity.

- **Sparsity of Labels:**

  how labels gathered in each movie are sparse.

  Eg: *label only scenes that seemed worthy of labeling (65.9% unlabeled).*

  Resolution: careful adjustment of the SVM's weighting.

Evaluation

Metrics: compute overlaps between two sets of items.
Two sets*: Human Selected set, Pulse Selected set.*

$$Precision = \frac{|\{\text{Human Selected} \cap \text{Pulse Selected}\}|}{|\{\text{Pulse Selected}\}|}$$

$$Recall = \frac{|\{\text{Human Selected} \cap \text{Pulse Selected}\}|}{|\{\text{Human Selected}\}|}$$

$$Fall - out = \frac{|\{\text{Non-Relevant} \cap \text{Pulse Selected}\}|}{|\{\text{Non-Relevant}\}|}$$

Evaluation expectation:

Higher values of precision and recall are better; the converse for fallout.

Summary of Results

- **Performance of Segment Rating**

Predicted segment ratings closely follow users' segment ratings*: average error of 0.7 on a 5-point scale; 40% improvement*

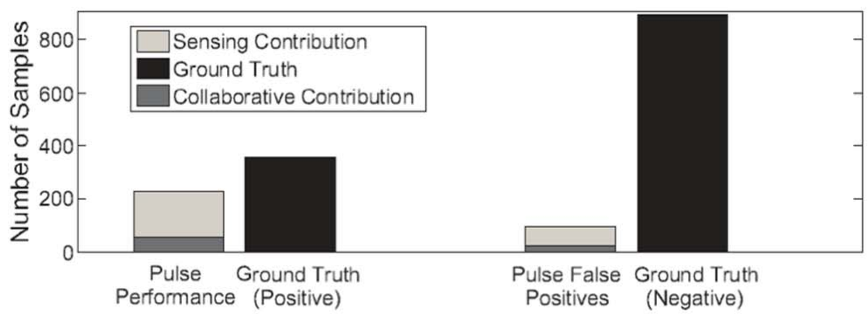


Figure 17. Break-up of contributions.

the contribution from sensing is substantial

Summary of Results (Cont.)

- **Performance of Final “Star” Rating**

Generates final ratings by thresholding the mean scores of per-minute segment ratings.
Result: Average error of 0.46 in the 5 point scale.

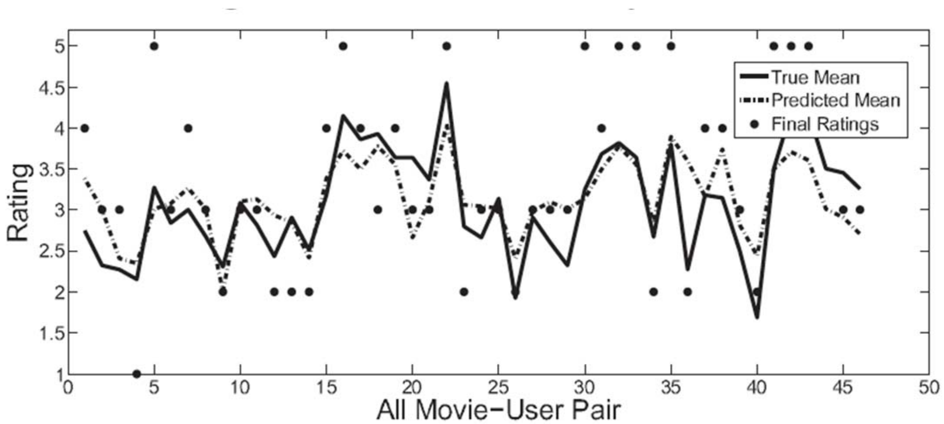


Figure 18. (a) Mean segment ratings and corresponding users’ final ratings.

Truth \ Pulse	1	2	3	4	5
1	0	1	0	0	0
2	0	4	2	0	1
3	0	1	17	0	1
4	0	0	2	5	2
5	0	0	2	1	7

(b) Confusion matrix.

(a) Users conservative on movie segments while generous on final rating.

(b) Higher values concentrate around the diagonal, indicating desired performance.
May have over-fitted data with thresholds.

Summary of Results (Cont.)

- **Performance of Label Quality**

  reaction labels (ground truth)
  & perception labels

Table 1. Label Vocabulary

Label Category	Vocabulary
Perception	Funny, Intense, Warm
Reaction	Laugh, Smile, Shaking, Focused, Distracted, Speaking

- Reaction Label Quality: great

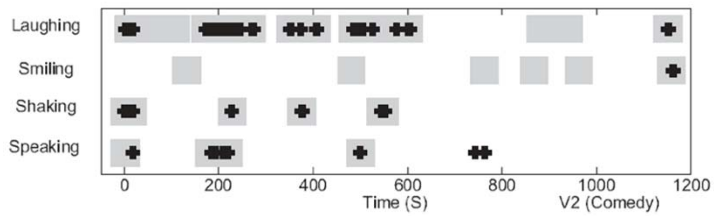


Figure 19. Reaction label prediction vs. groundtruth

- Perception Label Quality: weak

Reason: (1) their corresponding behaviors can be subtle and implicit; (2) users provided these labels for few segments.

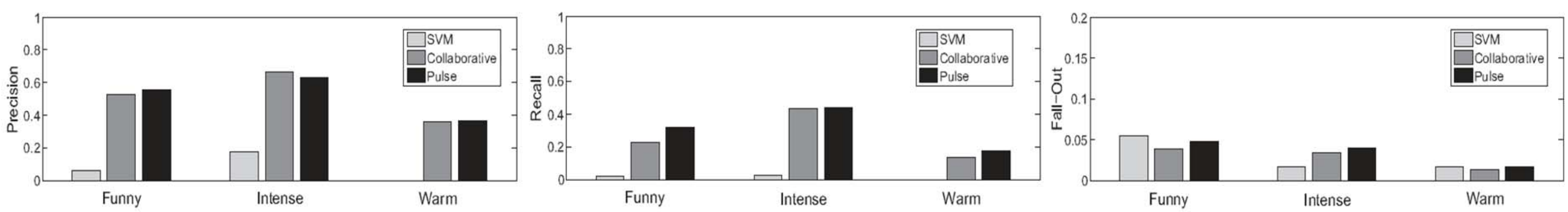


Figure 21. Performance comparison between SVM, collaborative filtering and our method (Pulse).

Summary of Results (Cont.)

- Tag Cloud and User Feedback
  - combine perception and reaction labels, each weighted by its normalized occurrence frequency.
  - “very cool”, “certainly useful information with zero extra burden”, “a richer tag set is needed”

- Power Consumption

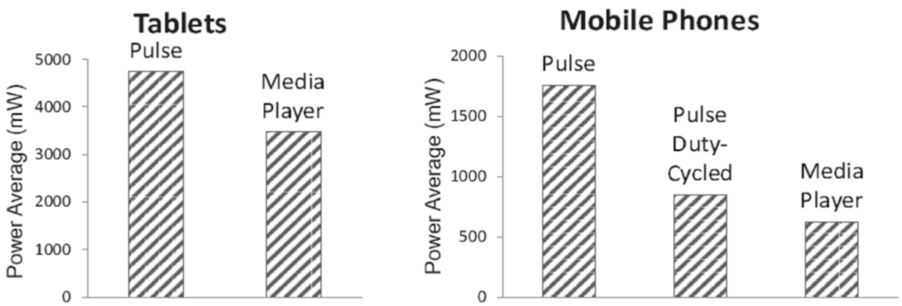


Figure 22. Power consumption comparison.

Related Work

- Activity Inference:

  a rich sensing platform; machine learning and inference;
  cause substantial power drain;
  efforts such as Little Rock could offload sensing to DSP chips, allowing the CPU to sleep

- Multimedia Annotation:

  TagSense: using sensor data from multiple devices, to annotate images

Conclusion & Future Work

- Use personal sensing and machine learning to build an application that automatically rates content on behalf of human users.
- Core idea: leverage device sensors to sense qualitative human reactions while she is watching a video; learn how these qualitative reactions translate to a quantitative value; and visualize these learnings in an easy-to-read format.
- On the technical side, the current label vocabulary is still limited;
  On the social side, may raise privacy concerns especially for exporting information to the cloud.

References

- 1. L. Bao et al. Activity recognition from user-annotated acceleration data. Pervasive Computing, 2004.
- 2. X. Bao and R. Roy Choudhury. Movi: mobile phone based video highlights via collaborative sensing. In Proceedings of the 8th international conference on Mobile systems, applications, and services. ACM, 2010.
- 3. H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. Computer Vision–ECCV 2006, 2006.
- 4. I. Cantador, M. Fern«andez, D. Vallet, P. Castells, J. Picault, and M. Ribi`ere. A multi-purpose ontology-based approach for personalised content filtering and retrieval. Advances in Semantic Media Adaptation and Personalization, pages 25–51, 2008.

References

- 5. M. Cherubini, R. De Oliveira, N. Oliver, and C. Ferran. Gaze and gestures in telepresence: multimodality, embodiment, and roles of collaboration. 2010.
- 6. E. Cuervo, A. Balasubramanian, et al. MAUI: Making Smartphones Last Longer with Code Offload. In ACM MobiSys, 2010.
- 7. P. Ekman and W. Friesen. Unmasking the face: A guide to recognizing emotions from facial clues. 2003.
- 8. R. Honicky, E. Brewer, E. Paulos, and R. White. N-smarts: networked suite of mobile atmospheric real-time sensors. In ACM NSDR, 2008.