Searching for Emotional Content in Digital Video

Aleksandra Sarcevic **Michael E. Lesk**

School of Communication, Information and Library Studies

Rutgers, The State University of New Jersey

4 Huntington Street, New Brunswick, NJ 08901

{aleksarc, lesk}@scils.rutgers.edu

ABSTRACT

Texts often contain high-level descriptive words, while a picture of a smiling face rarely includes the letters "happy." Annotating images with labels of this sort is extremely tedious and expensive. Thus, to explore more complex thoughts, information retrieval needs to move beyond term searching. Our interest is in multimedia, particularly video. We want to search videos for emotional content, but we lack the tools to do this kind of retrieval. To lay the groundwork for such tools, we investigated the ways in which people might best specify emotions they seek. In this position statement, we emphasize our work on identifying the most effective user interface techniques for formulating queries in video searches. Results from our exploratory studies can suggest the development of a tool for searching emotional content in videos, and contribute to a discussion of facial information processing and its relevance to HCI.

Keywords

Emotions, emotion recognition, facial expressions, information retrieval, digital video

INTRODUCTION

Retrieval of emotional content from multimedia databases has received little attention partly because search engines require explicit lists of facial features that typical users find hard to describe. This is unfortunate since the rise of digital video is about to flood computers with amateur and un-cataloged video recordings, mostly of people. We will want to search videos for emotional content, but we lack the tools to do this kind of retrieval.

Retrieval systems in general are acquiring some very specific kinds of feature analysis. The most familiar one is geographic: most Google users have probably noticed that typing something that looks like a street address gives a map of the location as an answer.

We anticipate that emotional labeling will be another kind of specific searching, particularly as people deal with un-cataloged video. Video is expensive to classify and index and manual annotation will only be feasible for a few professional productions. The vast bulk of home-made video is going to be un-cataloged, and is likely to have a great many scenes of people. Face identification is obviously going to be important, but other image attributes may be needed for accurate retrieval.

Although there is work on recognizing emotions in speech and photographs, it tends to be statistically based and not to provide a vocabulary in which newer emotions can be expressed. Our long term goal is not only to improve the ability to search for a few emotions, but to develop a system that can expand its recognition ability as users describe new emotions they might wish to search for.

MIT researchers [6] studied recognition of eight emotions—neutral, anger, hate, grief, platonic love, romantic love, joy, and reverence—using blood pressure, jaw clenching force, skin conductivity and breathing rate. Needless to say, these data are not available for video searching, and in addition, the limitation to an exact list of eight emotions makes it difficult for someone who wants to search for "fear" or "disgust."

Some earlier projects [1,5,7] have done more direct recognition of features such as "surprise" by explaining it as "raising brows and vertical expansion of mouth" but again, we would like to make an extensible version of such a recognizer, which can be taught by a user to expand its vocabulary and recognize additional emotions.

Studies by Ekman and his colleagues [2] have devoted considerable effort to analyzing facial expression through their Facial Action Coding Scheme (FACS). It is clear from their work that recognizing facial expressions based on component movements of at least 60 of the most important facial muscles requires specialized training. The utilization of FACS may well serve psychologists who analyze videotaped sessions for micro-expressions; however, it is not feasible to expect an average, "untrained" Internet user to understand the meaning of component movements of facial muscles.

We need to allow people to easily describe emotions they seek. However, those descriptions must be such that the

computers can automatically extract relevant facial features. Our goal is to identify the user interface techniques that best meet these requirements.

Since people are good at visual recognition, instead of asking them to break down emotions into constituent facial features, we believe that better approach is to let them create holistic visual representations of emotional expressions. We plan to achieve this by investigating the following aspects of different graphical tools: (1) the time it takes to create facial expressions; (2) how satisfied the user is with the created expression; and, (3) how well this visual description of emotion can be translated automatically into a computer query.

The user studies we report have provided information on (a) the types of facial features that typical users understand as signaling emotional content; and, (b) the effectiveness of potential input techniques for formulating queries. We briefly describe the design of our studies. Following this, we discuss our findings and conclusions.

USER STUDIES

To experiment with techniques for specifying emotions, we selected fifteen basic emotions listed in [4]: anger, fear, sadness, sensory pleasure, amusement, satisfaction, contentment, excitement, disgust, contempt, pride, shame, guilt, embarrassment, and relief. This expanded list of emotions provided enough space for experimenting and helped us determine the limits of a typical user's ability to distinguish subtle distinctions, such as those between "satisfaction" and "contentment."

To find out how well the typical users verbalize emotional expressions, we asked our participants to describe fifteen emotions in terms of observable facial features (e.g., open mouth, squinted eyes, wrinkled forehead, and so on). In the following sets of experiments, we asked people (a) to act emotions out, and (b) to draw them using a facial sketching program.

To measure how successful users are in specifying emotions by using these techniques, we presented the created sketches, video recordings of acting, and still frames extracted from the videos to the same and different participants and looked at their accuracy in labeling the emotions. Our participants (24 in total) were mostly graduate students recruited from the University.

RESULTS AND DISCUSSION

When asked to say which facial features they would associate with each emotion, our participants agreed on only few simple emotions (mostly on the happy—sad axis). For example, most of the participants suggested that an angry person has a red and a tense face. This proves to be a task people find difficult. It requires users to imagine faces, analyze them, and find proper words to accurately define an emotional expression. This may represent a significant challenge for users and increase their cognitive load. The results from this experiment confirmed our suspicion that users will not be good at verbalizing emotional expressions. Moreover, in their descriptions, participants often mentioned facial features that are not amenable to computer feature detection (e.g., facial color or muscle tone).

The quantitative analysis of similarity between the created sketches showed that participants could agree on certain, more distinct emotions (e.g., anger, excitement) and that the emotions that seem harder to differentiate (e.g., satisfaction, contentment, sensory pleasure) often had little agreement as to the specific facial properties needed. The users' primary efforts were spent on manipulating the mouth and eyes; the eyebrows and the nose seemed much less important.

Although participants made mistakes in identifying emotions from videos, in general they correctly differentiated between positive and negative emotions. We analyzed the correlations between different emotions as a way of determining how frequently people confused them and thus how similar they seemed to be. This is presented in Figure 1: the columns show the emotions expressed in video clips, and the rows on the left represent the labels attached to them. The number of instances of each assignment is given by hash marks and the background color gets darker as the number increases. This "confusion matrix" clearly indicates the emotions that our participants could not identify accurately. For example, they misinterpreted embarrassment for "shame" and "guilt;" similarly, "shame" was often misinterpreted for "embarrassment," and "guilt" for "shame."

With respect to the effectiveness of sketching and acting as potential techniques for formulating queries in video searches, we conclude that neither of these methods seems to be effective in specifying subtle differences of similar emotions. People differ widely in their skills at both sketching and acting, and we did not find either of these techniques to be an unambiguous way of specifying emotions. In fact, some of our participants aren't even any good at recognizing their own acting. Conversely, our participants performed well in identifying emotions from videos and still images, which suggests that a method such as "query-by-example" may be a better candidate for the query input. More studies are needed to investigate this important issue. It is necessary to allow users to specify queries both quickly and reliably. For, if the system is not fast, the users will grow impatient and annoyed with it. And, if the specified features are not reliable, the system will not yield satisfactory search results.

There are limitations in acting emotions and these are well known in psychology. People's expressions when they pretend feeling vs. when they genuinely feel emotions may differ. Therefore, acting emotions out may not be feasible technique; typical users will not be able to voluntarily move some muscles that are normally involved in expressing emotions [3].

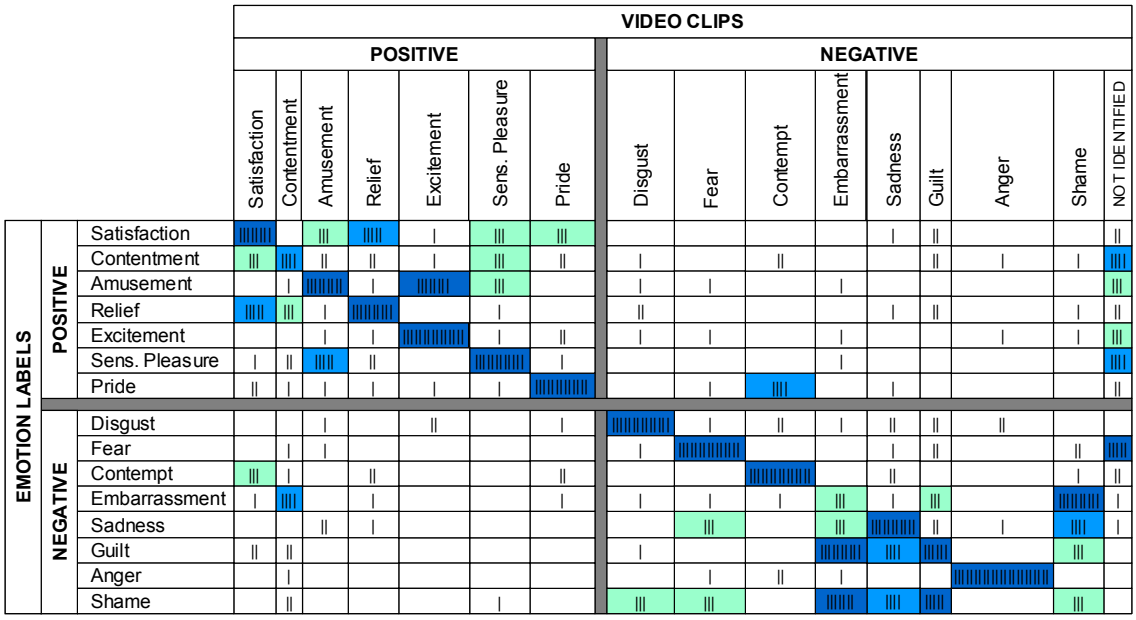

Figure 1. Confusion matrix for judgments across all participants and video sets. Each emotion was viewed on 28 instances—the maximum possible number of instances in each cell. The far right column ("not identified") represents the number of instances that users could not identify emotion from a video clip.

In short, our experiments indicate that:

- typical users are only reliable at distinguishing a small number of emotions, such as happy-sad or excited-placid;
- recognition rates for identifying emotions from videos are higher than for still images (about 30%) and sketches (about 50%); this suggests that users perform better at recognizing emotions in videos than in still images or sketches;
- there is an indication of cultural differences in the way facial features are understood, although we have insufficient data to describe this with confidence.

Simple sketches of faces may be preferable to video recordings in terms of feasibility of automatic extraction of facial features from users' descriptions. Although participants performed better in recognizing videos of people, automatic extraction of facial features may be difficult; simple sketches of faces may be better because the geometry of facial features can be automatically extracted from the user's input and used for querying. The fact that analysis of eye and mouth configuration is more valuable than of secondary or difficult to recognize features (e.g., muscle relaxation or facial color) may prove useful in automatic analysis of users' descriptions.

CONCLUSIONS

One important finding is that our users had difficulty both specifying and identifying subtle differences of similar emotions.

People are most reliable at distinguishing positive and negative emotions, i.e., between emotions such as "pleasure," "amusement," and "satisfaction" on one side and "anger," "disgust," and "shame" on the other side. Less certain as a second axis would be the difference between relatively excited states and relatively placid states, i.e., between labels such as "relief" and "sadness" compared with "excitement" and "anger." Additionally, our participants found many of the emotion labels confusing. The difference between emotional states such as "satisfaction" and "contentment" is difficult to explain even for a native English speaker, let alone to some of our participants who learned English as teenagers or adults. These results have implications for user interface design: developing a system that can distinguish between emotions that typical users cannot differentiate themselves may not be useful for searching.

Our findings from recognizing emotions in videos are independent of how powerful the software and user interface eventually might be because some of our users acted emotions and others performed recognition. Even those who looked at their own videos had problems identifying types of emotions they acted. Since people will remain much better in recognizing emotions than machines, this has some implications for determining the limits of a search engine and how accurate machine recognition should be. Although "better" results would be achieved with videos showing trained actors, we are not interested in them since such situations would be representative neither of typical video collections found on the Web, nor of users posing queries. However, it would be interesting to see

whether the emotions acted out by non-professionals would be more accurately recognized by trained people (e.g., certified user of FACS).

FUTURE RESEARCH

For future research, we plan to run extensive studies to determine the merits of the above techniques for specifying emotions. We will also explore other query formulating techniques. Since it appears that people are better at recognizing emotions in videos, this suggests that research on search engines might focus on tracking of facial features across multiple video frames, to help recognize dynamic features as well as static features. We also suggest that more exploration of cultural differences would be useful, given the world-wide use of Internet search systems.

ACKNOWLEDGMENTS

This research was supported by NSF Contract No. 0441172, and in part by donation from Siemens Corporate Research, Princeton, NJ.

REFERENCES

1. Black, M.J. and Yacoob, Y. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25, 1 (Oct. 1997), 23-48.
2. Ekman, P. and Friesen, W. V. *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press, 1978.
3. Ekman, P. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*, New York: Norton, 1995.
4. Ekman, P. Basic emotions, in T. Dalgleish and M. Power (Eds.). *Handbook of Cognition and Emotion*, Sussex, U.K.: John Wiley & Sons, Ltd., 1999.
5. Essa, I.A. and Pentland, A. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19, 7 (July 1997), 757-763.
6. Picard, R.W., Vyzas, E., and Healey, J. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23, 10 (Oct. 2001), 1175-1191.
7. Yacoob, Y. and Davis, L. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18, 6 (June 1996), 636-642.