Collective indexing of emotions in videos

Kathrin Knautz and Wolfgang G. Stock

Department of Information Science, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

Abstract

Purpose – The object of this empirical research study is emotion, as depicted and aroused in videos. This paper seeks to answer the questions: Are users able to index such emotions consistently? Are the users’ votes usable for emotional video retrieval?

Design/methodology/approach – The authors worked with a controlled vocabulary for nine basic emotions (love, happiness, fun, surprise, desire, sadness, anger, disgust and fear), a slide control for adjusting the emotions’ intensity, and the approach of broad folksonomies. Different users tagged the same videos. The test persons had the task of indexing the emotions of 20 videos (reprocessed clips from YouTube). The authors distinguished between emotions which were depicted in the video and those that were evoked in the user. Data were received from 776 participants and a total of 279,360 slide control values were analyzed.

Findings – The consistency of the users’ votes is very high; the tag distributions for the particular videos’ emotions are stable. The final shape of the distributions will be reached by the tagging activities of only very few users (less than 100). By applying the approach of power tags it is possible to separate the pivotal emotions of every document – if indeed there is any feeling at all.

Originality/value – This paper is one of the first steps in the new research area of emotional information retrieval (EmIR). To the authors’ knowledge, it is the first research project into the collective indexing of emotions in videos.

Keywords Emotion, Video, Collective indexing, Emotional information retrieval, Broad folksonomy, YouTube, Emotional intelligence, Classification

Paper type Research paper

1. Introduction

Can retrieval systems be used to search for feelings that are either expressed in documents or aroused in the recipient? We will apply the approach of Affective Computing (Picard, 1995) to the area of Human-Computer Interaction (HCI) (Boehner et al., 2007), and to research into information retrieval (Stock, 2007a), respectively. How can we store such emotive reactions so that they will be searchable and retrievable in the first place? Which indexing methods can be used? Are assessments of emotions consistent enough to be conclusively allocated to documents? These are core questions of the young research field of Emotional Information Retrieval (EmIR) (Schmidt and Stock, 2009). Emotions resonate in several document types. We can find “emotive” aspects in texts (e.g. novels or poems), music, images, videos and multimedia documents (e.g. weblogs).
In this article, we will report on the indexing of basic emotions in videos, as there are already results for music (Lee and Neal, 2007) as well as for images (Schmidt and Stock, 2009). We will try to bring the following research questions (RQ) closer to a first set of answers:

**RQ1.** Is a satisfactory retrieval for films on the Web practically realizable at the moment?

**RQ2.** What are basic emotions?

**RQ3.** How can we index basic emotions in order to apply them in practice in emotional retrieval?

**RQ4.** How consistent are user-oriented procedures for tagging emotions in videos?

**RQ5.** How many different users are needed for a stable distribution of the emotion tags? Are there stable distributions of the emotion tags at all?

**RQ6.** Can Power Tags corresponding to the displayed or felt emotions be derived from the tag distributions?

**RQ7.** Do certain emotions correlate with each other?

Our method was a survey of test subjects. A total of 776 persons watched 20 films each (prepared YouTube documents) in early summer of 2009 and tagged them via slide control according to nine different basic emotions.

### 2. Video retrieval (research question 1)

How can videos be made accessible for retrieval systems in the first place? Retrieval research roughly distinguishes between two approaches:

1. Concept-based information retrieval (via concepts and knowledge organization systems);
2. Content-based information retrieval (only via the content of the document).

A time- and personnel-independent approach is the automatic, content-based indexing of videos (Aigrain et al., 1996; Colombo et al., 1999; Del Bimbo, 1999; Enser, 2008a; Enser, 2008b; Gupta and Jain, 1997). In order to make content-based retrieval (Stock, 2007a, Ch. 31) possible, an analysis must first be performed, preparing the video for further steps. A film consists of many different scenes, which can be divided into shots. The shots, in turn, consist of single frames that represent the smallest video units (Smeaton, 2007).

These are visual components, summarized under Low-Level features. In general, Low-Level features describe metadata (Smeaton et al., 2008; Petrelli and Auld, 2008) and elementary video attributes that are composed of visual, auditive and textual components (Smeaton, 2007). The information that can be gleaned from these Low-Level features is limited, though, and can thus be used for an analysis of the significant and emotional content only with reservations. However, there are already approaches that attempt to automatically extract emotion-describing content from the Low-Level features (Chan and Jones, 2005; Chen et al., 2008; Hanjalic and Xu, 2005; Salway and Graham, 2003; Soleymani et al., 2008; Xu et al., 2005). Colombo et al. (1999, p. 41) point out that “the mapping of low-level color primitives into emotions is quite complex.”
Content-based video retrieval overall is still in its development phase, even though there is noticeable progress with regard to the automatic extraction of Low-Level features (Chan and Jones, 2005; Xu et al., 2005). But according to Smeaton (2007), further focal points must be set for the tasks. The determining of single features must be improved, as the results, particularly in the automatic filtering out of semantic properties, are still too inaccurate. Additionally, previous approaches in video retrieval are based mainly on keyframe and image retrieval (Petrelli and Auld, 2008; Kim and Kim, 2009), even though this static approach does not correspond to the desired video retrieval as the temporal aspect is missing, i.e. the movement of the objects and of the camera. Because of this, keyframes are only of limited usefulness as short forms of videos, as the selection process for them starts mostly in the middle of a shot and not in the place best suited to represent the video’s content (Kim and Kim, 2009). Furthermore, Smeaton (2007; Smeaton et al., 2008) criticizes the fact that at the moment, only single video shots are of determining value, which can hardly satisfy users searching for videos. A shot cannot really communicate semantic content, such as the mood and message of a video, which is why a scene made up of significant shots, reflecting the “true” content, should be created instead in order to meet the users’ demands.

Content-based video retrieval in general and the retrieval via emotions in particular carry with them a good few challenges yet. It would be important to determine what specific content (let’s say: a smiling face; Bailenson et al., 2008) arouses certain emotions (in most cases: happiness) in the viewer. Research on such “emotional points of reference” (relating to the “cognitive reference points” of Eleanor Rosch (1975)) remains to be performed.

In the following, concept-based video retrieval will be considered. Snoek and Worring (2009) discovered that in this area, concept-based solutions are often viewed as promising alternatives. In concept-based video retrieval, we work with concepts that could, in principle, also be automatically derived from the content (particularly from spoken text), but are, at the moment, mainly allocated intellectually (Schweins, 1997). A first way leads via knowledge organization systems (such as thesauri) and professional indexing. This route is taken, for example, by German and Austrian broadcasters in indexing TV productions. In case of intellectual indexing, classifications and a thesaurus are being used in indexing of television productions (ARD/ORF/ZDF, 2008) but feelings are not described. If a concept system for emotions were available, it would be possible, at least in principle, to intellectually allocate a controlled vocabulary on feelings to the videos (Kim and Park, 2003). This method is heavily dependent on the person of the indexer, though. A central problem of all content-based video retrieval is probably the indexing consistency, which is not very pronounced at all for images, and likely for films, as well (Markey, 1984). “The output of the indexing process seemed to be quite inconsistent” (Markkula and Sormunen, 2000, p. 273). From Goodrum (2000, p. 64), we learn that “manual indexing suffers from low term agreement across indexers (…), and between indexers and user queries.” “There is evidence that current systems for image access often fail the user,” Jørgensen (1998, p. 162) reminds us. So far, there have been no experiences with the intellectual indexing of emotions. “The influence of the photograph’s emotional tone on categorization has not been discussed much in previous studies;” Laine-Hernandez and Westman (2006) report. In the end, such an enterprise in the world wide web would run against a wall, as it is simply impossible (because unpayable) to have the millions of available videos analyzed by professional
indexers, because “manual assignment of textual attributes is both time consuming and costly” (Goodrum, 2000, p. 64). To put it bluntly: this ain’t the way (Enser et al., 2007).

The smaller time and personnel effort would seem to speak in favor of user-generated tags – i.e. social tagging or cooperative indexing (Jørgensen, 2007), as it can be found on video platforms such as YouTube (Geisler and Burns, 2007; Melenhorst et al., 2008; Van Velsen and Melenhorst, 2009). Folksonomies (Peters, 2009; Smith, 2008; Weller et al., 2010) are either “narrow” (only the author of the document may tag it; example: YouTube), “extended narrow” (apart from the author, his or her friends may give out tags, but only once per tag; example: Flickr) or “broad” (every user of the information service may allocate tags, which can consequently also be used multiple times for a single document; example: Del.icio.us). In narrow and extended narrow folksonomies, the problem of indexing consistency arises – in a particularly pointed manner, even, as it is not information professionals who index but laymen (Peters and Stock, 2007; Peters and Stock, 2008). In broad folksonomies and in case of a sufficiently large number of tagging users, there should be no consistency problems for the most commonly used tags; however, the tags are not from a standardized vocabulary but spring directly from the users’ language. Since YouTube uses a narrow folksonomy, we cannot use the indexing performed there for our purposes. In this way, Research Question 1 ends with a negative result: neither content- nor concept-based video retrieval leads to practically useful results for films on the WWW.

3. Basic emotions (research question 2)
In order to explain what “emotions” are, we will draw on results and discussions from psychology. There is a long tradition of research on emotions, and yet it has not been possible so far to produce a unified, exact definition for the concept of emotion. A definition is difficult in so far as everybody knows what an emotion is, but will find it hard to come up with a specific definition (Fehr and Russell, 1984). Schmidt-Atzert (1996, p. 18) also notes that so far there is no consensus on what to understand in an emotion. According to Izard (1994, p. 20), a complete definition of “emotion” must be comprised of three aspects:

1. the experience or conscious sensation of feeling;
2. the processes that take place in the brain and the nervous system; and
3. observable behavior, particularly in terms of facial expressions.

Kleinginna and Kleinginna (1981) analyzed 100 statements and definitions from pertinent specialist literature, dictionaries and introductory texts as early as 1981, and tried to come up with a preliminary definition: emotion is a complex model of changes, which comprises physiological arousal, feelings, cognitive processes and forms of behavior occurring in reaction to a situation which an individual has registered as significant for their person. Meyer et al. (2001) also conduct their research on the basis of a preliminary definition, as the definition itself represents a central question in research on emotion, and thus rather “is not the precondition of scientific research on emotion but its result” (Meyer et al., 2001, p. 50). Nevertheless, according to Meyer et al. (2001, p. 24), emotion can be characterized a little more exactly on the basis of various properties. Emotions are current psychic conditions of human beings, and can be delimited from other concepts such as dispositions and moods. Emotions are further
directed onto a specific, not necessarily existing object or objects that are the cause of these emotions. A further property is defined by the authors as the fact that emotions manifest themselves in the so-called reaction triad of subjective, behavioral and physiological aspects. In addition, emotions have a certain quality, intensity and duration. The quality serves as a group property, so that when an emotion occurs in a situation, such as for example “Person A is happy”, we speak of an allocation of the quality type happiness. The quality types’ concrete realizations distinguish themselves, besides a temporal limit, by a varying intensity. According to Meyer et al. (2001, pp. 29-30), the strength of an emotion can be determined on a scale from low via medium up to high intensity.

The question of how emotions come to be is being investigated by different theories of research on emotion, with various orientations. Behaviorist approaches stress the aspect of behavior and the conditions it is triggered by. Emotions are viewed as either innate (fear, anger and love) or are developed as conditioned reaction models to certain stimuli (Watson, 1930). Furthermore, there are evolution-psychological approaches, which emphasize the adaptive functions of emotions as conditioned by evolution, such as for example their survival function. These approaches go back to Darwin (1872), whose main objective was to prove the phylogenetical development of emotions, and have been continued by various emotion researchers (McDougall, 1926; Plutchik, 1980; Izard, 1971, 1977; Ekman and Friesen, 1971; Tomkins, 1962, 1963). Cognitive-physiological theories postulate that emotions are determined via the interaction between physiological changes. One develops and adapts his emotions via the (in)direct experience of these changes (James, 1884; Panksepp, 1982; Frijda, 1986). Attribution or attributional theories describe how one tries to understand and control the environment via cause attribution. According to these approaches, emotions are reactions to the results of actions (Arnold, 1960; Weiner and Graham, 1984).

In the course of our research, the question arises of which emotions should be selected for the project and consequently with regard to emotional video retrieval. In psychology, we find the approach of reducing emotions to a small but relatively fixed number with regard to a categorization of emotions. These are described as fundamental or basic emotions (Ortony and Turner, 1990). Due to the emotion researchers’ different theory orientations, though, there is no agreement as to their number.

Emotional music retrieval (Lee and Neal, 2007) as well as emotional image retrieval (Schmidt and Stock, 2009) work with five basic emotions: sadness, happiness, anger, fear and disgust. Considering the literature from psychology, we enriched the number of fundamental emotions we used for the project, which now are: sadness, anger, fear, disgust, surprise, desire, happiness and love. Independently of these basic emotions postulated in emotion psychology, a further aspect is taken on board in the form of fun, to do justice to this medial component as well. In accordance with the study of Shaver et al. (1987) fun cannot be regarded as a basic emotion, but many web queries are about opinions and qualities like “funny”, although people neglect to tag them (Bischoff et al., 2010).

4. Emotion-tagging via slide control (research question 3)
The answer to Research Question 1 leads – when positively formulated – to the discovery that the most useful method of indexing videos on the WWW is to use a broad folksonomy, i.e. to let a multitude of different persons index the same documents.
Research Question 2 leads us to accept nine basic emotions. Furthermore, we know that emotions have intensities. Building on these three results, in indexing instances of feeling we came to work

(1) with a controlled vocabulary for our nine basic emotions;
(2) with slide controls to adjust the intensity per emotion (Scale: 0 [non-existent] to 10 [very strong]);
(3) with the approach of the broad folksonomy, thus letting different users tag the documents via the different slide control settings.

We follow the approach of Lee and Neal (2007) for emotional music retrieval and of Schmidt and Stock (2009) for emotional image retrieval. Building on an idea in Schmidt and Stock (2009), we differentiate the emotions into represented emotions and emotions aroused in the viewer. After all, these do not have to match. Let us imagine, for example: a smiling hooligan (smiling being an expression of happiness) beating on rivaling soccer fans will cause most viewers not to smile in return, but rather to feel anger or disgust. The users also have the option of determining that the video does not express any emotion, or that they have no feelings in watching it. Additionally, we asked the viewers to state (in a sort of comment field) what they regard as the reasons for the emotions and their intensities. Figure 1 shows the slide control tagging of basic emotions we used.
This project does not engage with the distinction that some documents are constructed by the creator to be “easy to understand” and that other documents are created in such a way that the construction of meaning is located in the reader/viewers. The different levels of meaning and of consistency in the indexing are another question of interest (Rafferty, 2009).

5. Method: online survey on revised YouTube videos
The methodical processing of the research project comprises the following five steps:

1. Selecting and processing videos as test documents.
2. Pretest.
3. Selecting and writing to the test subjects.
4. Saving the subjects’ answers.
5. Analyzing the raw data.

Selecting and processing videos as test documents
All in all, 20 YouTube videos were chosen: two each that the project leaders identified as containing one of the nine basic emotions and two videos in which there was no representation of emotions, based on the interpretation of the two project leaders. Of the emotionally paired clips, one is underscored with music, the other is not. This was our way of preserving the option of analyzing whether music has any influence on the ascribed/observed emotions. Analysis of this data is pending. (A further project step, which we will also not address here, addresses 20 other films – in black and white this time. We hope to gain insights into emotional tagging for color/black-and-white films).

With such a large amount of videos for the test subjects to view and tag, the clips’ length plays an important role in selecting and processing them. We had to find a film length that would be enough, on the one hand, to carry the addressed emotions, and, conversely, as short as possible so as not to exceed the subjects’ attention span. In estimating the appropriate clip duration, we relied on the takeGas study by DoubleClick & Tomorrow Focus AG (Rohde and Lustig, 2007) as well as the study by Tubemogul (Burch, 2008). Both studies investigated the average length of people’s observation of online videos. The results diverge significantly. While the takeGas study shows an exit rate of 45 percent for durations longer than 20 seconds, the Tubemogul study arrived at a similar value, but only for durations longer than 60 seconds. As both studies differ in terms of context as well as the content displayed in the videos, we can only use the values as a point of orientation. Our decision was to concentrate on videos no longer than 25 seconds. Editing the videos, we used the video editing program Jashaka, which can be used to cut the clips to length and convert them to a suitable format.

Pretest
In a pretest preceding the project, the subjects, 14 students of information science, were able to alert us to contradictions, in terms of format and content. With a few corrections, we were able to optimize the test procedure.

Selecting and writing to the test subjects
The selection of the test subjects was performed via the mailing list of the computer center of the Heinrich-Heine-University Düsseldorf. The students on the list received an
e-mail alerting them to the research project and asking them to participate in the survey. For copyright reasons, the survey was protected by a password contained in the e-mail, excluding any users not affiliated with the university.

**Saving the subjects’ answers**
In order to the storage of the research data, and thus the analysis, the Heinrich-Heine University provided web space on a web server with PHP support and a MySQL database. A PHP application was then developed that dynamically created the survey’s HTML sites and saved the users’ survey data in the database. The graphical and functional properties of the survey were deposited within the PHP application. The single survey sites were based on a generic template, i.e. only the information for the respective video to be played was changed. The number of videos could be changed at will without accessing the application’s source code. At the beginning of the survey, each student was assigned a unique SessionID, saved in the participant’s browser via a cookie. This SessionID allowed the PHP application to identify the user on each site of the survey. The cookie was deleted after closing the browser.

The main components of the database are three charts. The first chart, *emo_user*, registered the participants’ statements regarding their department and gender. The respective SessionID and starting time was also stored in this part of the database. The video ratings entered via slide control were entered into the charts *emo_data* and *emo_data_complete*. When a complete rating of all 20 videos was accomplished, the data would be stored in the chart *emo_data_complete*. The chart *emo_data*, on the other hand, stored the ratings submitted in interrupted sessions, allowing us to analyze these data additionally. In order to access to the complete datasets, and thus the analysis, the data were transferred to an independent, password-protected site with an export function.

A total 776 subjects tagged all 20 videos. Of them 36 percent were male, 64 percent female. We had a total of 279,360 slide control adjustments (776 people * 20 clips * 18 emotions) as a basis for further calculations. We chose not to calculate significance values, as (nearly) all results lead to statistically significant results in the case of such large numbers. For our purposes, values of descriptive statistics were totally sufficient as, after all, we only wanted to find out whether consistent indexing of emotions is possible at all. We are aware of the fact that – with students of a single university – we have neither a random nor a representative, target-audience-specific sample. The large number of values allowed us to hope, nevertheless, to at least be able to trace the tendencies of emotional indexing.

**Analyzing the raw data**
We were able to use both Excel and SPSS in order to evaluate the raw data, as both applications facilitated the analysis of the data with regard to the formulated research questions.

**6. Consistency of emotional indexing (research question 4)**
The standard deviation is one of the most important distribution parameters of a discrete random variable. It is defined analogously to the respective empirical measures, and presupposes metrically scaled random variables. We calculated the standard deviation of the votes for all basic emotions (displayed and experienced) per
video, and then determined the arithmetic method of the standard deviations per emotion. The lower the value, the more consistently the subjects voted.

In total, both the expressed emotions (Table I) and the experienced emotions (Table II) showed relatively low values of standard deviation, and consequently high consistency values for the test subjects’ estimations. There is a very high consistency for the emotion of love in particular (with a median standard deviation of 0.88 for the expressed emotion and 0.66 for the experience), but disgust, sadness and anger are apparently also highly consent-inducing. Surprise surprises, as this emotion is (with a value of 1.94 for the expressed emotion) the most contentious of all in the voting. Experienced fun also leads to large differences in interpretation (the standard deviation being 1.75). These results reflect that fun occupies a special position. Fun is not a basic emotion and has been included in this study based on the search behavior of users. The results show that this emotion is seen very differently by the users.

The low values of standard deviations require an explanation. Often, only a few emotions dominate the videos, most of the times even just a single emotion. This is grounded in the selection of the clips, which have consciously been chosen based on emotional criteria. This leads to a very few basic emotions attaining a high intensity value, respectively, and the others a value of near zero. As the test subjects consistently

<table>
<thead>
<tr>
<th>Expressed basic emotion</th>
<th>Standard deviation</th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Love</td>
<td>0.88</td>
<td>0.94</td>
<td>0.82</td>
</tr>
<tr>
<td>Disgust</td>
<td>0.92</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>Sadness</td>
<td>1.11</td>
<td>1.09</td>
<td>1.12</td>
</tr>
<tr>
<td>Anger</td>
<td>1.21</td>
<td>1.11</td>
<td>1.30</td>
</tr>
<tr>
<td>Fun</td>
<td>1.46</td>
<td>1.48</td>
<td>1.44</td>
</tr>
<tr>
<td>Happiness</td>
<td>1.54</td>
<td>1.56</td>
<td>1.52</td>
</tr>
<tr>
<td>Desire</td>
<td>1.55</td>
<td>1.69</td>
<td>1.42</td>
</tr>
<tr>
<td>Fear</td>
<td>1.61</td>
<td>1.69</td>
<td>1.54</td>
</tr>
<tr>
<td>Surprise</td>
<td>1.94</td>
<td>2.15</td>
<td>1.74</td>
</tr>
</tbody>
</table>

**Notes:** $n = 776$ test subjects; 279,360 slide control adjustments; Question: What kind of emotion is depicted in the video?

<table>
<thead>
<tr>
<th>Experienced basic emotion</th>
<th>Standard deviation</th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Love</td>
<td>0.66</td>
<td>0.72</td>
<td>0.63</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.91</td>
<td>0.90</td>
<td>0.93</td>
</tr>
<tr>
<td>Anger</td>
<td>1.15</td>
<td>1.16</td>
<td>1.14</td>
</tr>
<tr>
<td>Desire</td>
<td>1.15</td>
<td>1.23</td>
<td>1.07</td>
</tr>
<tr>
<td>Disgust</td>
<td>1.24</td>
<td>1.28</td>
<td>1.19</td>
</tr>
<tr>
<td>Fear</td>
<td>1.36</td>
<td>1.37</td>
<td>1.34</td>
</tr>
<tr>
<td>Happiness</td>
<td>1.56</td>
<td>1.66</td>
<td>1.47</td>
</tr>
<tr>
<td>Surprise</td>
<td>1.71</td>
<td>1.77</td>
<td>1.65</td>
</tr>
<tr>
<td>Fun</td>
<td>1.75</td>
<td>2.01</td>
<td>1.63</td>
</tr>
</tbody>
</table>

**Notes:** $n = 776$ test subjects; 279,360 slide control adjustments. Question: What kind of emotion do you feel when you watch the video?
voted around zero, not only the arithmetic mean, but also the standard deviations have a value of near zero.

There is a side result that must not go unmentioned here. Separating the subjects according to gender provides some interesting insights: women tended to agree far less often on the expressed emotion than men – they only rated sadness, anger and surprise more consistently. There are also large differences for desire and surprise, the men being more consistent by far in this case. More differences are found for anger, which is ascribed to the videos more consistently by women. Experienced emotion is split along gender lines as well: men experience desire, happiness and joy far more consistently, for example. In the investigation of emotional tagging for images by Schmidt and Stock (2009, p. 871), a similar picture arose: there, women indexed less consistently than men across all emotions. An explanation for this observation has not been reached yet.

We can answer Research Question 4 in the positive: our test subjects did indeed consistently index emotions in videos. This means that a central hurdle with regard to emotional retrieval has been taken. We can assume that users (at least certain, delimitable user groups) allocate more or less the same basic emotions to the documents, with at least similar intensity. Whether this state of affairs is globally valid is yet to be answered. It might be possible, for instance, that users from different cultural circles might vote consistently with their peers, but choose different emotions across cultures.

7. Stable distribution of emotion tags (research question 5)
Halpin et al. (2007) used their study to measure the dynamics of tag distributions in folksonomies. Here they tried to determine at which time a stable distribution of tags is established in folksonomies. The measure they chose was months. With regard to collective tagging, this form of study is quite purposeful, as users of folksonomies continually tag documents in social-media platforms. In our research project, on the other hand, the subjects viewed and indexed all videos they were presented not over a period of months, but in a very short time. The goal was to find out which number of users would lead to a stable distribution form. Afterwards, the number of slide control adjustments might change, but without any repercussions on the shape of the distribution curve.

In order to find out how many participants are needed to provide for stability, the numbers of participants for the respective videos are divided into eight blocks with \( n \) participants: 1st Block: \( n = 76 \), 2nd Block: \( n = 176 \) etc. up to the 8th Block: \( n = 776 \). For each of these blocks, the average slide control adjustment per emotion was calculated for each video. We will exemplify this procedure with the concrete example of video no. 20 (Table III). The average slide control adjustment for anger (as expressed basic emotion) was 6.3 for the first (randomly selected) 76 participants, 6.6 for 176 subjects and so on, until the final value of 6.6 was reached with the number of 776 participants. The same procedure was repeated for all other emotions.

The evaluations make it clear that a stable distribution is reached as early as in the first block. Votes that break away from the arithmetic mean are “corrected” by the sheer number of users. This means: in emotional indexing practice, we do not require a large number of tagging users; if anything, the distribution of basic emotions “stands” with a few dozen users. There are two different forms of distribution that can appear: the Power Law and the inverse-logistical distribution. Both are typical curve
progressions for rankings of informational objects (Stock, 2006). In this study we identify these two distributions: if there are Power-Law distributions, one emotion dominates all others; if there are inverse-logistical distributions, several items heavily are represented. Sometimes users find no predominant emotions at all, which is resulting in no distribution.

8. Power tags for emotions (research question 6)
Can we conclusively separate from among the distributions those basic emotions that provide the best emotional description of the video in question? To solve this problem, we draw on the concept of “Power Tags” (Peters, 2009, pp. 363-372; Peters and Stock, 2007; Peters and Stock, 2008; Stock, 2007b; Stock and Stock, 2008, p. 172; Peters and Stock, 2010). “Power tags are tags that best describe the resource’s content, or the platform’s focal point of interest, according to Collective Intelligence (...), since they reflect the implicit consensus of the user community”, Peters defines (2009, p. 363). We are interested only in Power Tags on the resource level. Peters (2009, p. 370) emphasizes that the restriction of a search to Power Tags will increase that search’s precision, as the document-specific “Long Tail” of tags will no longer be searched.

In concurrence with Schmidt and Stock (2009, p. 871) we work with a threshold value for recording a basic emotion as a candidate for a Power Tag larger than or equaling 4. For all videos that do not have an emotion indexed with a slide control adjustment of at least 4, “no emotion” will be noted. Independently of the shape of the distribution, a statement on an emotion is only noted if the basic emotion has an intensity of at least 4. For the expressed emotions, no strong emotion is tagged in the case of six videos, for the experienced ones, the test subjects abstained in 11 cases.

As we have only nine basic emotions in total, it appears to make sense to regard all terms from the second-ranked one on down as the “Long Tail”, and to cut them off accordingly, in case of a Power Law distribution. Such a distribution follows the formula $f(x) = C/x^a$, where $x$ is the rank position, $C$ a constant and $a$ a value between, roughly, 1 and 2. To make sure, we will assume a small value for $a$ (at the moment, we calculate with 1). A curve is a Power Law precisely when the value of the item ranked 2nd is equal to or less than half of the value of the item ranked 1st. In all other cases, the form of distribution is noted as “inverse-logistical” (this does not correspond to the prevalent opinion in the literature – which defines the turning point of the distribution as a threshold value (Peters, 2009, p. 369), but is easily applicable). For inverse-logistical distributions, all emotions are noted whose intensity value is larger than/equal to 4. Thus it can definitely happen (as for the expressed basic

<table>
<thead>
<tr>
<th>Block</th>
<th>Number of participants ($n$)</th>
<th>Average slide control adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$n = 76$</td>
<td>6.3</td>
</tr>
<tr>
<td>2</td>
<td>$n = 176$</td>
<td>6.6</td>
</tr>
<tr>
<td>3</td>
<td>$n = 276$</td>
<td>6.8</td>
</tr>
<tr>
<td>4</td>
<td>$n = 376$</td>
<td>6.7</td>
</tr>
<tr>
<td>5</td>
<td>$n = 476$</td>
<td>6.7</td>
</tr>
<tr>
<td>6</td>
<td>$n = 576$</td>
<td>6.7</td>
</tr>
<tr>
<td>7</td>
<td>$n = 676$</td>
<td>6.6</td>
</tr>
<tr>
<td>8</td>
<td>$n = 776$</td>
<td>6.6</td>
</tr>
</tbody>
</table>

Table III. Average slide control adjustments for video No. 20 (expressed emotions) for the emotion of anger with growing numbers of indexers
emotions of Video 3) that three emotions are indexed. All emotions stated in Tables IV and V are being saved in the data fields for expressed or experienced basic emotions in the data sets of the video documents, and are available in a search for emotions.

The standard deviations in the rating of both the expressed and the experienced top emotions are relatively high – at least in comparison with the median values in Tables II and III. Such as result hints at the participants not being of one mind exactly as to how highly each dominant emotion should be rated in actuality, even though there is a general consensus on which emotion predominates (and only this observation is important for our purposes). Where a value of, say, 6 expresses high intensity for one user, another will set the slide control to 10, stating the same thing.

9. Correlations between the basic emotions (research question 7)
Which basic emotions co-occur? Are there typical constellations of feeling? To answer these questions, we calculate the correlation coefficient (per person, two-tailed) for all observed emotions.

Are emotions contagious (Wild et al., 2001)? If so, the emotions expressed in the videos would have to be experienced by the test subjects. For the Power Tags, there are indeed broad correlations between expressed and experienced emotions: in Video 1, it is fun, in Videos 3 and 5 it is surprise and fun, in 6 it is disgust, in 13 it is sadness, and in Video 19 it is fear. The correlation between the same feelings, i.e. between the same basic emotions deemed to be “expressed” and “experienced”, is shown in Table VI.

<table>
<thead>
<tr>
<th>Video no.</th>
<th>Expressed emotion</th>
<th>Median value</th>
<th>Standard deviation</th>
<th>Form of distribution</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Surprise</td>
<td>4.5</td>
<td>3.60</td>
<td>Inverse-logistical</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>3</td>
<td>Surprise</td>
<td>7.7</td>
<td>3.26</td>
<td>Inverse-logistical</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>Fun</td>
<td>5.8</td>
<td>2.81</td>
<td></td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>Happiness</td>
<td>4.0</td>
<td>1.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Fear</td>
<td>5.1</td>
<td>2.25</td>
<td>Inverse-logistical</td>
<td>2.2</td>
</tr>
<tr>
<td>5</td>
<td>Fun</td>
<td>5.7</td>
<td>3.40</td>
<td>Inverse-logistical</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>Surprise</td>
<td>5.4</td>
<td>3.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Disgust</td>
<td>5.0</td>
<td>3.97</td>
<td>Power Law</td>
<td>3.1</td>
</tr>
<tr>
<td>7</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>8</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>9</td>
<td>Love</td>
<td>8.1</td>
<td>4.04</td>
<td>Power Law</td>
<td>5.5</td>
</tr>
<tr>
<td>10</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>11</td>
<td>Anger</td>
<td>8.9</td>
<td>3.94</td>
<td>Power Law</td>
<td>8.1</td>
</tr>
<tr>
<td>12</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>13</td>
<td>Sadness</td>
<td>7.4</td>
<td>3.17</td>
<td>Power Law</td>
<td>4.4</td>
</tr>
<tr>
<td>14</td>
<td>Fun</td>
<td>4.7</td>
<td>4.02</td>
<td>Power Law</td>
<td>2.5</td>
</tr>
<tr>
<td>15</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>16</td>
<td>Love</td>
<td>7.6</td>
<td>4.22</td>
<td>Power Law</td>
<td>4.7</td>
</tr>
<tr>
<td>17</td>
<td>Desire</td>
<td>6.2</td>
<td>4.75</td>
<td>Power Law</td>
<td>2.8</td>
</tr>
<tr>
<td>18</td>
<td>Happiness</td>
<td>6.2</td>
<td>3.83</td>
<td>Power Law</td>
<td>2.5</td>
</tr>
<tr>
<td>19</td>
<td>Fear</td>
<td>8.9</td>
<td>4.05</td>
<td>Power Law</td>
<td>5.3</td>
</tr>
<tr>
<td>20</td>
<td>Anger</td>
<td>6.6</td>
<td>3.35</td>
<td>Power Law</td>
<td>5.8</td>
</tr>
</tbody>
</table>

Table IV.
Power tags for expressed basic emotions (emotions depicted in the documents)

Notes: All emotions with an intensity > 4; distance: to the intensity of the next emotion; n = 776 test subjects (for each video)
Particularly in the case of sadness and love, these feelings indeed seem to be contagious: expressed sadness, or expressed love, correlate with experienced sadness/love in the opinion of a lot of viewers. However, this is also valid (because of the various zero votes) for the case that an unexpressed emotion correlates with the same inexperienced emotion. We should therefore not put overdue emphasis on the results in this section.

Of the many significant correlations (all with a probability of error of less than 0.01), we will address only the most conspicuous ones here. Regarding all the data sets,
expressed happiness and expressed fun, as well as experienced happiness and experienced fun, frequently co-occur: the correlation coefficients for these cases are +0.476 and +0.386. Both emotions seem to be deemed equally related by the subjects, which is why the slide controls were often adjusted very similarly.

Negative correlations are mostly between positive and negative emotions. Expressed fear is thus in opposition to expressed happiness (−0.16) and to expressed fun (−0.16), and experienced fear to expressed happiness (−0.15) and expressed fun (−0.15). Analogous pairs of opposites are happiness and sadness as well as fun and sadness.

The basic emotions desire and surprise show no sustained, clear correlations to positive or negative feelings. Surprise, however, correlates positively (even strongly so) with expressed (+0.35) and experienced fun (+0.32), and negatively with expressed love (−0.12) and expressed sadness (−0.08), for example. Desire correlates positively with experienced love, for instance, but negatively with (expressed and experienced) anger (Table VII).

Positive emotions often correlate positively with other positive emotions and negatively with negative ones. The same goes, conversely, for negative emotions. There are, however, some emotions like desire and surprise, which correlate heavily with both positive and negative emotions. Here it seems to depend on the context, which determines whether the feelings get a positive or a negative inflection.

10. Conclusion and outlook
In conclusion, we will answer our research questions.

- **RQ1: Is a satisfactory retrieval for videos on the web practically realizable at the moment?** Content-based video retrieval is still in its infancy and not gainfully usable in practice at this time. There are experiments to recognize content-based emotions in videos, but we are a long way away from satisfactory solutions. Concept-oriented video retrieval works either with a controlled vocabulary or with user-generated tags. If a single indexer tags a video, or if we are dealing with narrow/extended narrow folksonomies, the problem – massive in this case – of indexing inconsistency rears its head. To master the masses of videos on the web, only the use of folksonomies can be considered. The ideal way would be the use of broad folksonomies. These are not being used by currently active video services (e.g. YouTube), however. A satisfactory video retrieval is not to be found today. An emotional video retrieval with practical usefulness does not exist.

- **RQ2: What are basic emotions?** Preliminary investigations of emotional retrieval (Lee and Neal, 2007, and Schmidt and Stock, 2009) as well as the psychological literature on emotions suggest that there are at least five (happiness, sadness, fear, anger, disgust), but probably even nine (additionally: surprise, desire, love, fun) fundamental emotions. Emotions have an intensity, i.e. they can be experienced strongly or weakly.

- **RQ3: How can we index basic emotions in order to apply them in practice in emotional retrieval?** In order to rectify this unfortunate situation in video retrieval, we suggest the use of a controlled vocabulary (the concepts of the basic emotions) and slide controls (to evaluate the intensity of the emotion) in the context of a broad folksonomy for indexing the emotional content of videos.
<table>
<thead>
<tr>
<th></th>
<th>Love</th>
<th>Happiness</th>
<th>Fun</th>
<th>Surprise</th>
<th>Longing</th>
<th>Sadness</th>
<th>Anger</th>
<th>Disgust</th>
<th>Fear</th>
<th>Love</th>
<th>Happiness</th>
<th>Fun</th>
<th>Surprise</th>
<th>Longing</th>
<th>Sadness</th>
<th>Anger</th>
<th>Disgust</th>
<th>Fear</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expressed emotions (related to the documents)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Love</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happiness</td>
<td>0.382</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fun</td>
<td>-0.103</td>
<td>0.476</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surprise</td>
<td>0.013</td>
<td>0.230</td>
<td>0.347</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Longing</td>
<td>0.241</td>
<td>-0.036</td>
<td>-0.129</td>
<td>-0.063</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sadness</td>
<td>0.033</td>
<td>-0.140</td>
<td>-0.159</td>
<td>-0.117</td>
<td>0.339</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anger</td>
<td>-0.117</td>
<td>-0.111</td>
<td>-0.006</td>
<td>-0.071</td>
<td>-0.116</td>
<td>-0.089</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disgust</td>
<td>-0.100</td>
<td>-0.059</td>
<td>-0.046</td>
<td>0.177</td>
<td>-0.101</td>
<td>-0.099</td>
<td>-0.070</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fear</td>
<td>-0.094</td>
<td>-0.164</td>
<td>-0.157</td>
<td>0.135</td>
<td>0.019</td>
<td>0.056</td>
<td>-0.038</td>
<td>0.234</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Experienced emotions (related to the users)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Love</td>
<td>0.469</td>
<td>0.081</td>
<td>-0.009</td>
<td>-0.036</td>
<td>0.128</td>
<td>0.012</td>
<td>-0.057</td>
<td>-0.069</td>
<td>-0.065</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happiness</td>
<td>0.146</td>
<td>0.184</td>
<td>0.388</td>
<td>0.119</td>
<td>-0.022</td>
<td>-0.096</td>
<td>-0.067</td>
<td>-0.064</td>
<td>-0.106</td>
<td>0.151</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fun</td>
<td>-0.105</td>
<td>0.166</td>
<td>0.394</td>
<td>0.298</td>
<td>-0.130</td>
<td>-0.125</td>
<td>0.003</td>
<td>0.045</td>
<td>-0.078</td>
<td>0.060</td>
<td>0.386</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surprise</td>
<td>-0.119</td>
<td>0.075</td>
<td>0.232</td>
<td>0.331</td>
<td>-0.102</td>
<td>-0.079</td>
<td>-0.052</td>
<td>0.089</td>
<td>0.106</td>
<td>-0.070</td>
<td>0.157</td>
<td>0.317</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Longing</td>
<td>0.198</td>
<td>-0.005</td>
<td>-0.103</td>
<td>-0.078</td>
<td>0.242</td>
<td>0.130</td>
<td>-0.086</td>
<td>-0.078</td>
<td>-0.069</td>
<td>0.289</td>
<td>0.032</td>
<td>-0.093</td>
<td>-0.089</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sadness</td>
<td>0.009</td>
<td>-0.116</td>
<td>-0.132</td>
<td>-0.103</td>
<td>0.225</td>
<td>0.474</td>
<td>-0.072</td>
<td>-0.089</td>
<td>0.028</td>
<td>0.053</td>
<td>0.089</td>
<td>-0.122</td>
<td>-0.022</td>
<td>0.171</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anger</td>
<td>-0.067</td>
<td>-0.040</td>
<td>-0.022</td>
<td>-0.031</td>
<td>-0.058</td>
<td>-0.050</td>
<td>0.316</td>
<td>-0.040</td>
<td>-0.020</td>
<td>-0.050</td>
<td>-0.066</td>
<td>-0.066</td>
<td>-0.025</td>
<td>-0.044</td>
<td>-0.017</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disgust</td>
<td>-0.120</td>
<td>-0.118</td>
<td>-0.111</td>
<td>0.019</td>
<td>-0.133</td>
<td>-0.126</td>
<td>-0.069</td>
<td>0.293</td>
<td>0.075</td>
<td>-0.084</td>
<td>-0.101</td>
<td>-0.107</td>
<td>0.076</td>
<td>-0.096</td>
<td>-0.103</td>
<td>0.017</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Fear</td>
<td>-0.104</td>
<td>-0.145</td>
<td>-0.144</td>
<td>0.048</td>
<td>-0.095</td>
<td>0.051</td>
<td>0.022</td>
<td>0.383</td>
<td>-0.056</td>
<td>-0.107</td>
<td>-0.114</td>
<td>0.124</td>
<td>-0.002</td>
<td>0.028</td>
<td>0.240</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** \( n = 776 \) test subjects; 279,360 slide control adjustments; correlations per person (two-tailed); italics: contagious emotions; bold: high correlation
• RQ4: How consistent are user-oriented procedures for tagging emotions in videos? The consistency of users’ votes, measured via the standard deviation from the mean value, is high enough (roughly between 1 and 2 on a scale from 0 to 10) for us to assume a satisfactory consensus between the indexers. Some feelings – particularly love – are highly consent-inducing.

• RQ5: How many different users are needed for a stable distribution of the emotion tags? Are there stable distributions of the emotion tags at all? There are indeed stable distributions of emotions per video. These distributions become clear even with very few numbers of users (<100) and stay stable in shape, no matter how many further users tag the video. We can separate three typical forms of distribution: no emotion (and as a result no distribution), Power-Law distribution and inverse-logistical distribution.

• RQ6: Can Power Tags corresponding to the displayed or felt emotions be derived from the tag distributions? In the distribution of basic emotions according to a Power Law, exactly one emotion is filtered out, which is allocated to the video as an emotional point of access. In the case of inverse-logistical distributions, it is several emotions (two or three) that are attached to the video. We must distinguish between expressed emotions (related to the documents) and experienced emotions (related to the users’ feelings), as the users rate them differently as well.

• RQ7: Do certain emotions correlate with each other? For the correlations, we distinguish between the analysis of identical emotions as expressed and experienced, as well as the analysis of the correlations between different emotions. “Contagious” basic emotions (i.e. those with high correlation values between their expressed and experienced iterations) are largely sadness and fun. Positive correlations across emotions are particularly happiness and fun, negative ones are, for example, fear with happiness and fun.

• Outlook. Looking at the results of our large scale indexing experiment we can assume, with good reason, that users are able to recognize (preset) basic emotions in videos consistently. To maintain stable distributions of emotions per video, we require only relatively small numbers (<100) of participating users. These results are in line with the results for emotional image indexing (Schmidt and Stock, 2009). This encourages us to take our next step in the EmIR research program and to start developing a prototype of a retrieval system for emotional documents – called “Media Emotion Search (Memose)” (Knautz et al., 2010).

References


Jørgensen, C. (2007), “Image access, the semantic gap, and social tagging as a paradigm shift”, *Proceedings of the 18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research, Milwaukee, WI*.


Watson, J.B. (1930), *Behaviorism*, University of Chicago Press, Chicago, IL.


Further reading


About the authors

Kathrin Knautz is a Researcher at the Information Science Department of the Heinrich-Heine-University Düsseldorf (Germany); her main research topics are emotional information indexing and retrieval and Web 2.0 retrieval.

Wolfgang G. Stock is the Head of the Department of Information Science, Heinrich-Heine-University Düsseldorf; his research areas are information retrieval, knowledge representation, informetrics and the information market. Wolfgang G. Stock is the corresponding author and can be contacted at: stock@phil-fak.uni-duesseldorf.de

To purchase reprints of this article please e-mail: reprints@emeraldinsight.com
Or visit our web site for further details: www.emeraldinsight.com/reprints