IMPLICIT HUMAN-CENTERED TAGGING

*A.Vinciarelli, N.Suditu**

{vincia,nsuditu}@idiap.ch
Idiap Research Institute (CH)

M.Pantic†

m.pantic@imperial.ac.uk
Imperial College London (UK)
EEMCS - University of Twente (NL)

ABSTRACT

This paper provides a general introduction to the concept of Implicit Human-Centered Tagging (IHCT) - the automatic extraction of tags from nonverbal behavioral feedback of media users. The main idea behind IHCT is that nonverbal behaviors displayed when interacting with multimedia data (e.g., facial expressions, head nods, etc.) provide information useful for improving the tag sets associated with the data. As such behaviors are displayed naturally and spontaneously, no effort is required from the users, and this is why the resulting tagging process is said to be "implicit". Tags obtained through IHCT are expected to be more robust than tags associated with the data explicitly, at least in terms of: generality (they make sense to everybody) and statistical reliability (all tags will be sufficiently represented). The paper discusses these issues in detail and provides an overview of pioneering efforts in the field.

Index Terms— Implicit Tagging, Nonverbal Behavior Analysis

1. INTRODUCTION

This paper proposes the idea of using human behavior analysis techniques for *implicit tagging*, i.e., for tagging multimedia data independently of explicit tags associated with the data. In other words, implicit tagging means that a data item could get tagged each time a user interacts with it, based on the reactions of the user to the data (e.g., laughter when seeing a funny video), in contrast to explicit tagging paradigm in which a data item gets tagged only if a user actually decides to associate tags with it.

Tagging has emerged in the last years in *social media* sites where the users are not only passive consumers of data, but active participants in the process of creating, diffusing, sharing, and assessing the data delivered through websites [1]. These sites give the possibility to the users to add keywords (explicit tagging) to the data that are then used for indexing and retrieval purposes. Tagging represents a major novelty with respect to previous data retrieval approaches because, for the first time, the indexing stage (the representation of the data in terms suitable for the retrieval process) is not computing-centered, i.e., performed through a fully automatic process driven purely by technological criteria, but human-centered, i.e., performed through a collaborative effort of millions of users following the natural modes of social interaction over the network [2].

However, although tagging should represent a major step towards bridging the semantic gap, because taggers are expected to annotate the data in terms of how they perceive the content, the analysis of tagging behavior has clearly revealed that such an expectation is too optimistic [3]. The reason is that people are not driven by technological criteria, i.e., they do not aim at making retrieval systems to work, but by *personal* and *social* needs [2][3]. This gives rise to at least two major problems.

- *Egoistic tagging*. When taggers are driven by *personal* purposes and needs, they tend to use tags that are meaningless to other users like, e.g., *John-and-Mary-at-my-place*. These tags are unlikely to appear in queries submitted by any other user, thus they are not useful from a retrieval point of view.
- *Reputation driven tagging*. When taggers are motivated by *social* purposes, they tag large amounts of data to increase their reputation in the on-line communities formed around social networking sites. For example, as a result, their tags have a disproportionate influence on the retrieval process. In fact, as the occurrences of tags follow Zipf-like laws [4], a tag appearing just few tens of times ends up having large weight in any statistical retrieval approach, due to the fact that most of the tags occur less than half a dozen times in total.

These problems are aided and abetted by "fraudolent" behaviors like adding tags that have nothing to do with the content

*The work of A. Vinciarelli has been supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet). The work of N. Suditu has been supported by the Hasler Foundation through the EMMA project.

†The work of M. Pantic has been supported in part by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet), and in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

of the data, but bringing messages that the taggers want to convey (e.g., taggers can tag the data with their name to get known).

Extracting *effective* tags, i.e., tags oriented to aiding correct functioning of retrieval technologies, based on the spontaneous behavior of users is the core idea of Implict Human-Centered Tagging. Such tags could be added to the tag sets explicitely associated with the data and limit the effect of the above listed problems.

Research in psychology suggests that people behave with machines in the same way they behave with other people [5]. Exactly this fact, that they display their reactions in front of the computer, e.g., while interacting with multimedia data (e.g., shaking their head or frowling when encountering incorrectly tagged data) is the basis of implicit tagging paradigm. The analysis of behavior of users interacting with multimedia data could help to capture information useful for implicit tagging in terms of the following.

- *Assessing explicit tags*. Users retrieve data based on their tags. Reactions like surprise and disappointment when presented with retrieval results might mean that the tags associated with the data are incorrect (e.g., something *grewsome* is tagged as *funny*).
- *Assigning new explicit tags*. The user behavior might provide information about the content of the data. If the user laughs, the data can be tagged as *funny*, if the user shows disgust or revulsion, the data can be tagged as *horror*, etc.
- *User profiling*. The user behavior might reveal specific needs and attitudes of each user. For example, if the user squints each time the data from a specific website/datapool is retrieved, this might be a sign that the user has difficulties in viewing the data, which may result in flagging the data source as being less favourable for this user.

The rest of this paper presents an overview of previous work in HTC. Previous attempts of including the human in the retrieval process are discussed in Section 2. Kinds of tags that can be extracted from human behavior are discussed in Section 3. Section 4 concludes the paper.

2. HUMAN-CENTRIC APPROACH TO RETRIEVAL

Earliest works on Information Retrieval, dating back to the fifties, considered explicitly only approaches of fully computing-centric nature. Statements typical for these early works were in the following fashion: "*It should be emphasized that this system is based on the capabilities of machines, not of human beings*" [6]. In the sixties the paradigm shifted towards involving the human in the retrieval process. Even though people have been included in the process not as *humans*, but as *users*, i.e., the interaction was constrained to few "button-like" functions that retrieval systems provided (e.g., downloading a given document), this was still a major shift in paradigm towards human-centered approaches.

The most successful approach of this kind is *Relevance Feedback* (RF), which requires users to identify the most relevant documents among those retrieved by a system in response to an initial query Q. The documents identified by the users are then modeled to formulate a second query Q' that typically improves the retrieval results (see [7] and [8] for surveys on RF in text and image retrieval, respectively). Other approaches involving users in the retrieval process are *query log analysis* (e.g., see [9]), modeling of past documents seen by a given user (e.g., see [10]), and the large body of works dedicated to user adaptation (see [11] for a survey).

In the last ten years, automatic analysis of human behavior has been the subject of significant attention in the computing community (see [12] for an extensive survey). The most interesting aspects of this research, from an IHCT point of view, is the analysis of affective states (*affective computing* [12]) and social signals (*Social Signal Processing* [13]). These technologies can help realize automatic *user behavior* to *human behavior* modeling, and human-hehavior-based tagging and retrieval systems, bringing around long sought solution to flexibile yet general, non-tiresome yet statistically reliable, multimedia tagging and retrieval.

The multimedia retrieval community recognizes that this is needed [14], but, to the best of our knowledge, only few efforts have been made to include human behavior in the retrieval loop [15][16]. Furthermore, except of the works investigating the role of emotions in information seeking [15] and ranking [16] these works mostly try to understand the emotional content of the data, i.e., what emotions are displayed by people portrayed in the data (see e.g., [17]) or what emotions can be elicited by the data (see e.g., [16][18]), rather than the actual behavior of the users [14].

3. TAGGING BASED ON NONVERBAL BEHAVIOR

IHCT is an attempt to address the above-outlined gap and move a step further towards a Human-Centered approach, where one of the most natural modes of human communication, nonverbal behavior, is sensed and analyzed to enrich and improve the tag sets associated with the data. Nonverbal behavior typically occurs in human-human communication and conveys information about whatever cannot be said with words (e.g. emotions, feelings, attitudes, etc.) [13]. As already mentioned above, there is evidence that people display the same nonverbal behavior when interacting with computers as when interacting with other people [5]. This means that the analysis of nonverbal behavior in front of a computer can provide hints about the feelings, attitudes, and reactions of users with respect to the tagged data they interact with. This is potentially a major source of *effective* tags, i.e., tags that make sense to all users and are sufficiently represented

to allow reliable statistical modeling. There are several cues conveyed by nonverbal behavior, that could be used as tags for the data, namely emotional (affective) cues, level of interest, and focus of attention.

Automatic analysis of emotions has been extensively investigated. Proposed approaches rely mainly on recognition of facial expressions (see [12] for the most recent survey) and vocal behavior [19], but recent works suggest that also body gestures and combinations of different nonverbal cues should be taken into account as well [20]. Furthermore, emotions have been shown to play an important role in Human-Computer Interaction [21] and tagging can be considered as a way of interacting with multimedia data. From a tagging point of view, emotional signals (like laughter, frowns and head shakes in disagreement, nose wrinkling and horizontal mouth stretching in disgust, etc.) are interesting because their interpretation could be used as tags. These would be effective because they make sense to all users and, if there is only a limited number of those (like funny, horror, sad, etc.), they can be sufficiently represented to allow reliable statistical modeling.

Furthermore, emotional signals are often machine detectable since these involve reactions like laughing, sobbing, frowling, head nods and shakes, jaw drops, scretching, etc. Laughter detection has been addressed, e.g., in [22] through vocal behavior analysis and in [23] through combination of vocal and facial behavior. Also, detection of various gestures like facial gestures, head gestures, and hand gestures have been extensively researched in recent years (see [12][13] for survey papers in the field). Recently, few related works have been published investigating the role of emotions in information seeking [15], and ranking movie scenes based on user-affect-related physiological signals [16]. However, tagging multimedia data based on emotional signals have not been attempted yet, to the best of our knowledge.

Nonverbal behavior conveys information about how much people are interested in what happens around them, as well as on what attracts their attention. The interest level can be detected through facial expression analysis [24], body posture [25], and combination of vocal and facial behavioral cues [26]. Attention is mainly captured through gaze tracking (see [27] for a survey on gaze detection and tracking methods), and head pose recognition [28]. Both interest and attention can provide hints about how much the data are appreciated by users, and can lead to the attribution of tags like *thumbs-up* and *thumbs-down*, and can lead to development of recommendation mechanisms.

4. CONCLUSIONS

This paper introduces the concept of Implicit Human-Centered Tagging, where the basic idea is improving tag sets associated with multimedia data using the behavioral feedback of users. IHCT represents a research direction towards tagging systems that would rely on natural modes of human interaction and on technologies for automatic human behavior analysis, in particular on emotion recognition, interest level detection, and attention analysis. All these domains have been investigated more or less extensively in the recent years, but, to the best of our knowledge, they have never been applied for tagging purposes.

There are several challenges that the researchers face in the field of IHCT. Behavioral feedback is often culture dependent – in some cultures, it is usual to inhibit spontaneous reactions and reactions observed in one culture do not have to be the same to those observed in another culture for the same stimulus (e.g., a joke considered funny in one culture can be offensive in another one). Furthermore, human user's behavior is influenced not only by the data that he or she is interacting with, but also by other factors such as user personality (introvert persons are less likely to display their emotional reactions) and transient conditions like stress and fatigue that decrease the reactivity of users. Finally, most of the commercially available computers are equipped with microphones and cameras, but these sensors are not always of sufcient quality for conducting automatic human behavior analysis. However, the goal of IHCT is not to model reactions of each and every user, but to annotate the data with tags representing common users' reactions (e.g., "*amusing*", "*unpleasant*", etc., or in terms of valance and arousal). Although the tag collection will be limited to those users who show their reactions (as opposed to those who inhibit their reactions or are expressionless), who have appropriate equipment, etc., this will allow for filtering the noise from the tags because reactions determined by user specic conditions would not have a major statistical impact. Furthermore, although this paper has mainly discussed collection of behavioral feedback from audiovisual sensors, the value of the information that could be collected by using physiological sensors must not be underestimated. Changes in hartbeat, clamminess, respiration rate, etc., are reliable cues to detection of affective and mental states [29] and, as these signals cannot be consciously controlled, they could be extremely valuable for IHCT tools, especially in cases where spontaneous behaviors are inhibited for cultural or contextual factors (e.g., when the user is a library or another public space).

The development of IHCT systems represents not only a potential way of improving current tagging systems, but also a step towards human-centered approaches for Information Retrieval, a domain that so far has been characterized by mostly computing-centric approaches.

5. REFERENCES

[1] K. Lerman and L. Jones, "Social browsing on Flickr," in *Proc. Intl. Conf. on Weblogs and Social Media*, 2007.

[2] M. Ames and M. Naaman, "Why we tag: motivations for annotation in mobile and online media," in *Proc.*

SIGCHI Conf. on Human Factors in Computing Systems, 2007, pp. 971–980.

[3] O. Nov et al., "What drives content tagging: the case of photos on Flickr," in *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, 2008, pp. 1097–1100.

[4] C. Cattuto, V. Loreto, and L. Pietronero, "From the cover: Semiotic dynamics and collaborative tagging," *Proc. Natl. Ac. of Sci.*, vol. 104, no. 5, pp. 1461, 2007.

[5] C. Nass and S. Brave, *Wired for speech*, The MIT Press, 2005.

[6] M. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958.

[7] I. Ruthven et al., "A survey on the use of relevance feedback for information access systems," *The Knowledge Engineering Review*, vol. 18, no. 2, pp. 95–145, 2003.

[8] X.S. Zhou and T.S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Systems*, vol. 8, no. 6, pp. 536–544, 2003.

[9] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, "Analysis of a very large web search engine query log," in *ACM SIGIR Forum*, 1999, vol. 33, pp. 6–12.

[10] S. Dumais et al., "Stuff I've seen: a system for personal information retrieval and re-use," in *Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2003, pp. 72–79.

[11] A. Kobsa, "Generic user modeling systems," *User Modeling and User-Adapted Interaction*, vol. 11, no. 1, pp. 49–63, 2001.

[12] Z. Zeng, M. Pantic, G.I. Roisman, and T.H. Huang, "A survey of affect recognition methods: audio, visual and spontaneous expressions," *IEEE Trans. on Patt. An. and Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[13] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: survey of an emerging domain," *Image and Vision Computing, to appear*, 2009.

[14] M.S. Lew et al., "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. on Multimedia Computing, Communications, and Applications*, vol. 2, no. 1, pp. 1–19, 2006.

[15] I. Arapakis, J.M. Jose, and P.D. Gray, "Affective feedback: an investigation into the role of emotions in the information seeking process," in *Proc. ACM SIGIR Intl. Conf. Research and development in information retrieval*, 2008, pp. 395–402.

[16] M. Soleymani, G. Chanel, J.J.M. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *Proc. ACM Workshop Multimedia semantics*, 2008, pp. 32–39.

[17] A. Salway and M. Graham, "Extracting information about emotions in films," in *Proc. ACM Intl. Conf. on Multimedia*, 2003, pp. 299–302.

[18] A. Hanjalic and L.Q. Xu, "Affective video content representation and modeling," *IEEE Trans. on Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[19] K.R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[20] H. Gunes et al., "From the lab to the real world: Affect recognition using multiple cues and modalities," *Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition*, pp. 185–218, 2008.

[21] R. Cowie et al., "Emotion recognition in human-computer interaction," *IEEE Sig. Proc. Mag.*, vol. 18, no. 1, pp. 32–80, 2001.

[22] K.P. Truong and D.A. van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, no. 2, pp. 144–158, 2007.

[23] S. Petridis and M. Pantic, "Audiovisual laughter detection based on temporal features," in *Proc. IEEE Intl. Conf. on Multimodal Interfaces*, 2008, pp. 37–44.

[24] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *IEEE Trans. on Multimedia*, vol. 8, no. 3, pp. 500–508, 2006.

[25] S. Mota and R. Picard, "Automated posture analysis for detecting learners interest level," in *Proc. Workshop on Computer Vision and Pattern Recognition for Human Computer Interaction*, 2003.

[26] B. Schuller et al., "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, 2009.

[27] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 116–134, 2007.

[28] K. Smith, S.O. Ba, J.M. Odobez, and D. Gatica-Perez, "Tracking the visual focus of attention for a varying number of wandering people," *IEEE Trans. on Patt. An. and Mach. Intell.*, vol. 30, no. 7, pp. 1212–1229, 2008.

[29] J.T. Cacioppo et al., "The psychophysiology of emotion," in *Handbook of Emotions*, M. Lewis and J.M. Havil-Jones, Eds., pp. 173–191. Guilford, 2000.