IMPLICIT HUMAN-CENTERED TAGGING

Maja Pantic & Alessandro Vinciarelli

Tagging is the annotation of multimedia data with user-specified keywords known as tags, with the aim of facilitating fast and accurate data retrieval based on these tags. In contrast to this process, also referred to as Explicit Tagging, Implicit Human-Centered Tagging (IHCT) refers to exploiting the information on user's nonverbal reactions (e.g., facial expressions like smiles or head gestures like shakes) to multimedia data, with which he or she interacts, to assign new or improve the existing tags associated with the target data. Thus, implicit tagging allows that a data item gets 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 is requested (or chooses) to associate tags with it. As nonverbal reactions to observed multimedia are displayed naturally and spontaneously, no purposeful explicit action (effort) is required from the user; hence, the resulting tagging process is said to be "implicit" and "human-centered" (in contrast to being dictated by computer and being "computer-centered").

Tags obtained through IHCT are expected to be more robust than tags associated with the data explicitly, at least in terms of generality and statistical reliability. To wit, a number of human behaviors are universally displayed and perceived – e.g., basic emotions like happiness, disgust and fear – and these could be associated to IHCT tags such as "funny" and "horror", which would make sense to everybody (generality) and would be sufficiently represented (statistical reliability).

EXPLICIT TAGGING

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 Internet websites [7]. These sites allow users to assign keywords (explicit tags) 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 (i.e., the representation of the data in terms suitable for the retrieval process) is not computer-centered, that is, performed through a fully automatic process driven solely by technological criteria, but human-centered,

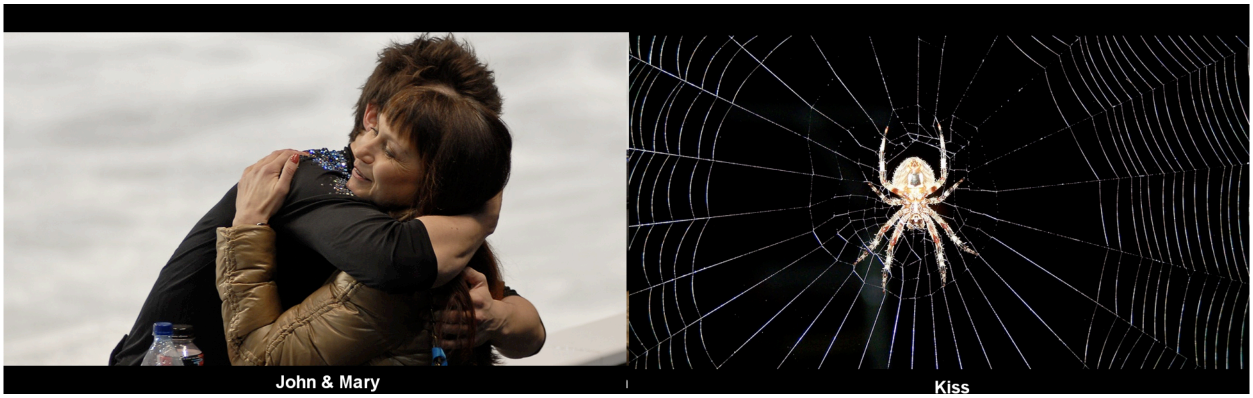

Figure 1: Tagging driven by personal needs or by individual interpretation of the content.

that is, performed through a collaborative effort of millions of users following the natural modes of social data sharing over the network [1].

However, in contrast to widely expected results, data retrieval approaches based on user-specified tags proved to be rather inaccurate in practice. The reason is that, when tagging, people are not driven by the aim of making retrieval systems work well, but by individual interpretation of the content, personal and social needs, and sometimes asocial behavior. In turn, this results in the following.

- *Egoistic tagging* – When users are driven by personal needs or by individual interpretation of the content, rather than by factual description of the content, they tend to use tags that are meaningless to other users like (for examples, see Figure 1). These tags will aid erroneous retrieval or will never appear in queries of other users and are, therefore, useless from a data retrieval point of view.
- *Reputation-driven tagging* – When users are motivated by “social” goals like reputation, they tag large amounts of data to increase their reputation in the on-line communities formed around social networking sites. Hence, as a result, their tags end up having a disproportionate influence on the retrieval process. More specifically, as the occurrences of tags follow Zipf-like laws, a tag appearing a few tens of times ends up having a large weight in any statistical retrieval approach due to the fact that most of the tags occur less than half a dozen of times in total.
- *Asocial tagging* – When users aim to put forward certain messages, they may tag large amounts of data with the target messages, which do not have anything to do with content of the data (e.g., users may tag the data with their name in order to get known, or with a certain parole like ‘666’ or ‘anarchy!’).

IMPLICIT TAGGING

Implicitly extracting effective tags, such that they aid accurate data retrieval and are based on spontaneous (non-elicited) nonverbal reactions of the user interacting with the target multimedia data (for examples of such reactions, see Figure 2), is the core idea of Implicit Human-Centered Tagging. Such tags could replace or complement the explicit tags that were associated with the data to limit the effect of the above listed problems. More specifically, implicit tagging could be used for the following purposes.

- *Assessing the correctness of explicit tags* – Users retrieve data based on their queries that are then matched against explicit tags associated with the data. Reactions like surprise and disappointment when presented with the retrieval results might mean that the tags associated with the data are incorrect, resulting in an inaccurate retrieval (e.g., something gruesome is tagged as funny). Associating an implicit tag indicating likelihood that the associated explicit tag is incorrect could facilitate lower ranking of the target datum next time when the same query is presented to the system.
- *Assigning new explicit tags* – The user's nonverbal reactions to multimedia data might provide information about the content of the data in question. 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's behavior and reactions to multimedia data might reveal specific needs and attitudes of each user. For example, if the user squints each time the data from a specific website / data pool 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 favorable for this user. Thus, an implicit tag could be associated with this data indicating that the user in question favors less this particular website, facilitating lower ranking of the target data next time when the system presents the results to this particular user.

Implicit, human-behavior-based tagging and retrieval systems could bring around a long-sought solution to flexible, general, non-tiresome, and statistically reliable multimedia tagging and retrieval. To the best of our knowledge and in spite of recognized need for such systems [4], only few efforts have been made so far to include the observed user's reactions and behavior into the retrieval loop (e.g., [6]).

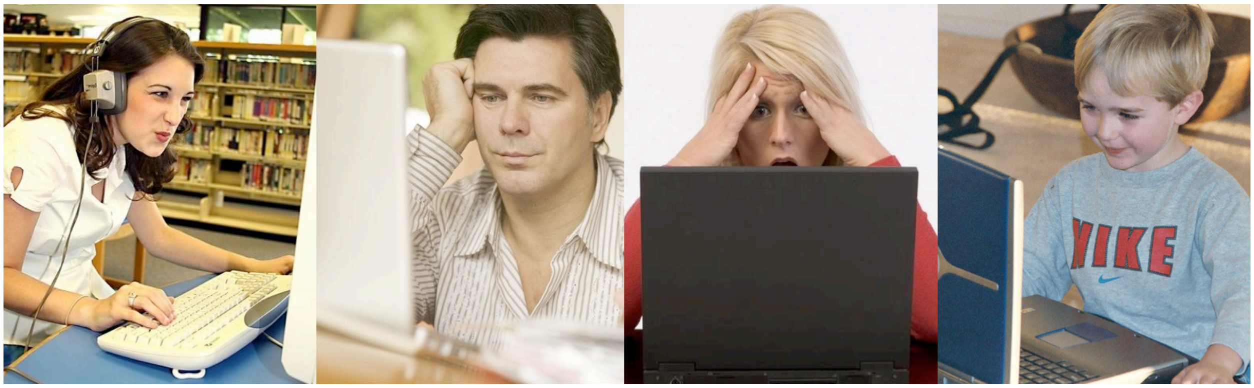

Figure 2: Examples of spontaneous (non-elicited) nonverbal reactions when interacting with multimedia data.

Two main problems impeding the progress in this field are: (i) that automatic analysis of human spontaneous reactions and behavior in front of the computer is far from being a trivial task, and (ii) that a proper inclusion of implicit tags in the data tagging and retrieval loop is yet to be investigated.

HUMAN BEHAVIOUR IN HUMAN-COMPUTER INTERACTION

Research on the border between human-computer interaction (HCI) and psychology emphasizes the phenomenon called “Media Equation”, [11] – people react to multimedia data (images, videos, audio clips) in the same way as they react to real objects and they interact and behave in front of a computer in the same way as they would interact with another person (except of speaking, which is less frequent in human-computer interaction, if present at all, due to the current computers’ inability to maintain lively and intelligent spoken dialogue for extensive periods of time). Hence, automatic analysis of the user’s nonverbal behavior conveyed by facial expressions, body gestures, and vocal outbursts like laughter (for examples, see Figure 2), which are our primary (and often unconscious) means to communicate affective, attitudinal and cognitive states [2], could provide valuable hints about the data that the user is currently involved with. Exactly this fact forms the basis of the implicit tagging paradigm.

Of course, not all human nonverbal behaviors are expected to be useful for data tagging and retrieval. Yet, behavioral cues revealing user’s affective states like amusement or revulsion, some cognitive processes like attention (interest) and boredom, and some attitudinal states like (dis)liking and (dis)agreement (e.g., with an existing explicit tag), could potentially be a major source of *effective tags*, that is, tags that make sense to everybody (generality) and are sufficiently represented (statistical reliability).

AUTOMATC ANALYSIS OF HUMAN AFFECT

Human natural affective behaviour is multimodal, subtle and complex. It is communicated multimodally by means of language, vocal intonation and vocal outbursts, facial expression, hand gesture, head movement, body movement and posture [2]. Yet, the mainstream research on automatic human affect recognition has mostly focused on either facial or vocal expressions analysis in terms of seven discrete, basic emotion categories (neutral, happiness, sadness, surprise, fear, anger and disgust, see Figure 3), and then based on data that has been posed on demand or acquired in laboratory settings [18].

Research findings in psychology indicate that in everyday interactions people exhibit non-basic, subtle and rather complex affective and cognitive states like thinking, interest or embarrassment, and that deliberately and spontaneously displayed behaviour have differences both in morphology of the display (i.e., which audio, visual and tactile cues have been displayed) and in its dynamics (i.e., which cue has been displayed when, how fast, for how long, etc.). Hence, as complex, natural displays of affective

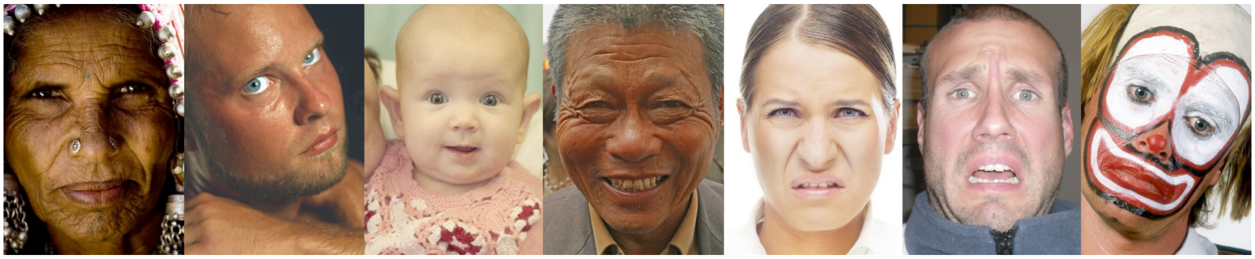

Figure 3: Facial expressions of basic emotions: neutral, anger, surprise, happiness, disgust, fear and sadness.

behaviour are conveyed via tens (or possibly hundreds) of anatomically possible facial expressions, vocal outbursts, bodily gestures and physiological signals, a single label (or any small number of discrete classes) may not reflect the complexity of the affective state conveyed by such rich sources of information. Hence, a research strand in psychology advocates the use of dimensional description of human affect, where an affective state is characterized in terms of a small number of latent dimensions such as valance (the degree of pleasantness) and arousal (the degree of excitement). Accordingly, the research in automatic human affect analysis has recently started to shift towards modelling, analysis and interpretation of the subtlety, complexity and continuity of naturalistic (rather than acted) affective behaviour in terms of latent dimensions, rather than in terms of a small number of discrete emotion categories [18], [5]. However, considering the fact that different affective states may have similar or identical valence or arousal values (see Figure 4), it remains unclear whether the dimensional approach to automatic interpretation of affective behaviour is the best approach or whether automatic affect analyzers should attempt to recognize distinct, non-basic emotion categories.

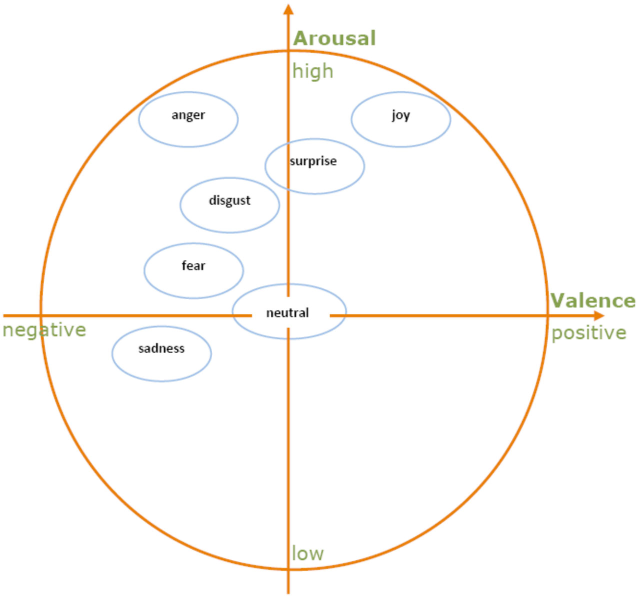

Figure 4: Mapping of basic emotions to valance-arousal space.

Progress in both directions has been recently reported. Several efforts have been reported on automatic analysis of spontaneously displayed facial and/or vocal affect data either in terms of non-basic affect categories like fatigue and pain [18], or in terms of latent dimensions [5]. For example, Wollmer et al. [17], proposed a novel method for continuous vocal affect recognition in terms of valance and arousal values. The method applies Long Short-Term Memory Recurrent Neural Networks to functionals of acoustic Low-Level Descriptors, representing the input features extracted from the whole utterance to be classified. It achieved an average recognition rate of 87% and 94% for valance and arousal, respectively, when trained and tested on a database of spontaneous vocal behavior exhibited in a simulated human-virtual-agent interaction scenario. This method is a pioneering effort in attaining automatic continuous analysis of human affect in terms of latent dimensions [5]

Also, few studies have been reported on automatic analysis of spontaneously produced affect data from multiple non-conventional modalities including body gestures and bio signals [5], and few studies investigated automatic, vision-based discrimination between spontaneous and deliberate affective behavior [9]. For example, Valstar et al. [15], proposed an automated system for distinguishing acted from spontaneous smiles. They have shown that combining information from multiple visual cues (in this case, facial expressions, head movements, and shoulder movements) outperforms single-cue approaches to the target problem. They used the motion of facial components (eyes, eyebrows, and mouth), head, and shoulders as input to a classifier combining ensemble and statistical learning (more specifically, Gentle Boost and Support Vector Machines) and achieved a recognition rate of over 93% for the target problem. The study clearly shows that the differences between spontaneous and deliberately displayed smiles are in the dynamics of shown behavior (e.g., the amount of head and shoulder movement, the speed of onset and offset of the actions, and the order and the timing of actions' occurrences) rather than in the configuration of the displayed expression, which is in contrast to other approaches to automatic discrimination between spontaneous and acted human behavior, which are typically based on morphological rather than on temporal differences in behavior [9].

Some of these efforts would be valuable for implicit tagging since recognized user's affective states like amusement (expressed in terms of latent dimensions as: positive valance, high arousal), disgust (expressed in terms of latent dimensions as: negative valance, high arousal), fear (expressed in terms of latent dimensions as: negative valance, neutral to high arousal), or surprise (expressed in terms of latent dimensions as: neutral valance, high arousal) could be used to assign new tags to the data with which the user interacts (e.g., funny, disgusting, horror, etc.), as well as to reason about the correctness of the existing explicit tag associated with the data (user's surprise might be an indication of incorrectly tagged data). Also, automatic analysis of whether the user shows spontaneous (genuine) affect or she acts it, could be valuable for implicit tagging as spontaneous smile would indicate amusement while acted (e.g., ironic) smile could be an indication of incorrectly tagged data. Such tags would be effective because they make sense to all and, if there were only a small number of these, they would be sufficiently represented to allow reliable statistical modeling. However, it is important to note that automatic analysis of naturalistic affective behavior in all its complexity and subtlety is just beginning to be investigated [18],

[5], and robust, reliable methods that could form the basis for inclusion of human affective behavior into the data tagging and retrieval loop are yet to be developed.

AUTOMATC ANALYSIS OF COGNITIVE PROCESSES AND ATTITUDINAL STATES

When it comes to cognitive processes like attention (interest) and boredom and attitudinal states like (dis)liking and (dis)agreement, very few efforts towards automatic recognition of these states have been reported so far [16]. Arguably the most advanced method proposed up to date for detecting the level of interest is that by Schuller et al. [13], who applied Support Vector Regression on a large number of features extracted from the audiovisual utterance to be classified including facial expressions, speech, acoustic features, and non-linguistic vocalisations like laughter and hesitation. The method applies previously reported techniques such as Active Appearance Models for facial expression recognition and Bag-of-Words for linguistic analysis, and achieves an average recognition rate of approximately 70% for continuous analysis of the level of interest in spontaneous behavioural data recorded in a face-to-face interview setup.

Both interest and agreement level can provide hints about how much the user appreciates the data retrieved based on the given query, and can be used for user profiling (e.g., if the interest level is low, an implicit tag could be associated with the retrieved data indicating that the user favors less the website in question) and for assessment of the correctness of the existing explicit tags (e.g., if the user shows signs of disagreement, an implicit tag indicating likelihood that the associated explicit tag is incorrect could be associated with the retrieved data). In turn, these implicit tags could be used to develop better data retrieval and recommendation mechanisms.

Attention (interest) level can be captured by means of gaze tracking (gaze aversion or staring at a single point are signs of inattentiveness), head pose estimation and tracking (this is an alternative to gaze tracking), facial expression analysis (dropped eyelids, frequent slow blinks, mouth corner dimpling, etc., are signs of fatigue and boredom), body posture analysis (supporting the head by a hand and unerect posture are signs of boredom), and vocal outbursts like yawning (a prominent signal of fatigue and boredom). Disliking and disagreement can be captured by means head gesture analysis (head shake and head nod are typical signs of disagreement and agreement), facial expression analysis (smirk, lip bite or wipe, lip puckering or tightening, nose flaring or wrinkling, etc., are all signs of disagreement), body gesture and posture analysis (arms folding and leaning back are signs of disagreement), hand gesture analysis (clenched fist, forefinger raising or wiggling, and hand wag, are typical signs of disagreement), and vocal outbursts analysis (while sighs and throat-clearing are typical signs of disagreement, laughter is a typical sign of agreement).

Although current efforts towards automatic analysis of interest and agreement level are mostly single-cue based, research in computer vision and signal processing has advanced significantly in the past years to allow fast and moderately accurate recognition of the above-mentioned visual and audiovisual behavioural cues, which allow development of multimodal multi-cue approaches.

AUTOMATC ANALYSIS OF BEHAVOURAL SIGNALS

Sensing human behavioral signals including facial expressions, body and hand gestures, and non-linguistic vocalizations, has witnessed a lot of progress in the past years.

To determine the direction of the gaze, eye tracking systems employ either the so-called red-eye effect, i.e., the difference in reflection between the cornea and the pupil, or computer vision techniques to find the eyes in the input image and then determine the orientation of the irises. There are now several companies that sell commercial eye trackers like Tobii, SMI GmbH (Figure 5), EyeLink, Interactive Minds, etc. Although realizing non-intrusive (non-wearable), robust, and accurate eye tracking remains a difficult problem, most of the commercially available eye trackers will work well in Human-Computer Interaction (HCI) scenarios like multimedia browsing, in which the user remains seated in front of the computer screen.

Common approaches to head pose estimation and tracking include appearance-based approaches (they match the input image of the head to previously stored examples), feature-based approaches (they use the location of facial features like nose, mouth, and eyes, to determine the head pose), manifold embedding methods (they seek low-dimensional manifolds that model the continuous variation in head pose), non-linear regression methods (they use a functional mapping from the image to a head pose measurement), non-rigid modeling approaches (they fit a personalized non-rigid model like Active Appearance Model or Elastic Bunch Graph to the facial structure in the image), tracking methods (they recover the global pose change of the head from the observed movement in the input video, see Figure 5 for an example), and hybrid methods that combine two or more of the above-mentioned methods [8]. Similarly to the state of the art in eye tracking, although fast, robust, and accurate head pose estimation and tracking remains a difficult problem in unconstrained environments, several existing methods will work well in HCI scenarios like multimedia browsing.

To facilitate detection of subtle facial signals like a frown or a smile, several research groups begun research on machine analysis of facial muscle actions (atomic facial signals also referred to as action units, AUs, [9]; e.g., AU4 relates to frowning, AU12 relates to smiling, AU18 relates to lip puckering, etc.). As AUs are independent of interpretation, they can be used for any higher-order decision making process including recognition of affective states, cognitive processes like attention (interest) and boredom, and attitudinal states like (dis)liking and (dis)agreement. A number of promising prototype systems have

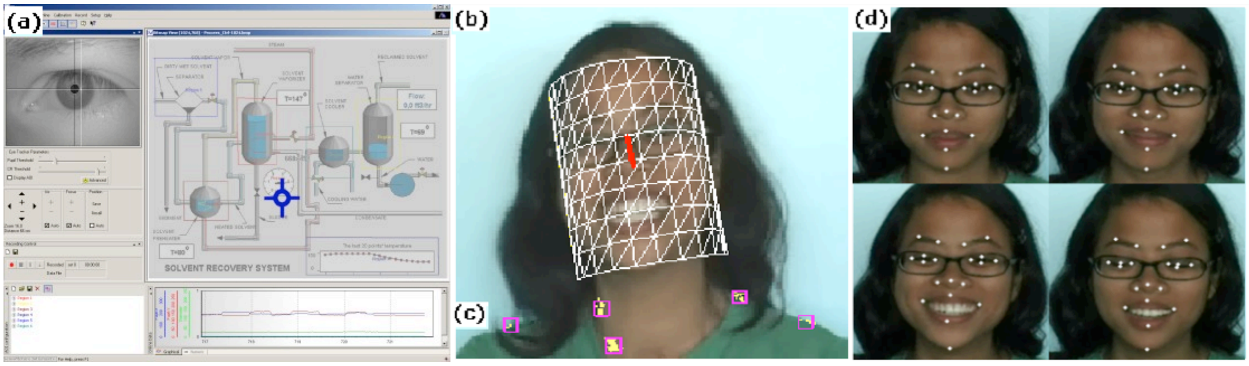

Figure 5: Examples of tools for (a) eye tracking (SMI GmbH), (b) head tracking (used in [15]), (c) shoulder movement tracking (used in [15]), and (d) facial point tracking used in [10], [15]).

been proposed that can recognize 15 to 27 AUs (from a total of 32 AUs) in either (near-) frontal view or profile view face image sequences depicting deliberately displayed facial behaviour [9]. Most of these employ statistical and ensemble learning techniques and are either feature-based (i.e., use geometric features like facial points or shapes of facial components, see Figure 5 for an example) or appearance-based (i.e., use texture of the facial skin including wrinkles, bulges, and furrows). One of the main criticisms that these works received is that the methods are not applicable in real-life situations, where subtle changes in facial expression typify naturalistic facial behavior rather than the exaggerated changes that typify deliberately displayed facial behavior. Hence, the focus of the research in the field started to shift to automatic AU recognition in spontaneous facial expressions (produced in a reflex-like manner). Several works have recently emerged on machine analysis of AUs in spontaneous facial expression data [9]. These methods use probabilistic, statistical, and ensemble learning techniques, and perform with reasonably high accuracy in more or less constrained environments (e.g., where no occlusion occurs and the variation in head pose and illumination is small). However, since the present systems for facial AU detection typically depend on accurate head, face, and facial feature tracking, they are still rather limited in performance and robustness when the input recordings are made in less constrained environments such as the multimedia browsing scenario, in which the user can turn the head away from the screen, occlude the face by hand, or work under natural light conditions which can change from moment to moment.

Because of its practical importance and relevance to human activity recognition and surveillance and sign language recognition, automatic analysis of body postures and hand and body gestures is nowadays one of the most active fields in computer vision. Common techniques include model-based methods (they use geometric primitives like cones and spheres to model head, trunk, limbs and fingers), appearance-based methods (they use color and/or texture information to track the body and its parts), salient-points-based methods (they use local signal complexity or extremes of changes in the entropy in space and time that correspond to peaks in hand or body activity variation), and spatio-temporal shapebased methods (they treat human body gestures as shapes in space-time domain). Most of these methods emphasize Gaussian models, probabilistic learning, and particle filtering framework. Under the assumption that the user's hands will always be visible and that he or she will not move the hands except to manipulate the mouse or to make a specific sign of boredom or disagreement (e.g., clench the fist, support the head with a hand, cross the arms, etc.), current methods could work reasonably well to facilitate recognition of attention and agreement level based on hand and body gestures. However, this assumption is rather unrealistic. In casual human behavior in front of the computer, the hands do not have to be always visible (under the table, on the back of the neck, and under the hair), they may be in a cross-fingered position, and one hand may be (partially) occluded by the other. Also, body and hands detection and tracking in unconstrained environments where large changes in illumination and cluttered or dynamic background may occur still pose significant research challenges. Although some progress has been made to tackle these problems using the knowledge on human kinematics, most of the present methods cannot handle such cases correctly.

Since research findings in psychology argue that listeners are rather accurate in decoding distress, anxiety, boredom, and sexual interest from nonlinguistic vocalizations like laughs, cries, sighs, coughs and yawns, few efforts towards automatic recognition of these nonlinguistic vocal outbursts have been recently reported. Most of these efforts are based only on audio signals. However, since it has been shown by several experimental studies in either psychology or signal processing that integrating the information from audio and video leads to an improved performance of human behavior recognition, few pioneering efforts towards audiovisual recognition of nonlinguistic vocal outbursts have been recently reported including audiovisual analysis of laughter [10]. These methods use probabilistic or statistical learning techniques, and are based on standard audio features like Mel-Frequency Cepstral Coefficients (MFCCs) or Perceptual Linear Predictive (PLP) coefficients and video features obtained through tracking facial components like mouth, eyes, and eyebrows. Although it is still unclear whether audio-based detectors of vocal outbursts can be used in real-world HCI scenarios like multimedia browsing, this goal seems to be reachable [12]. On the other hand, audiovisual detectors of vocal outbursts that can work in real-world scenarios are not available yet, mainly due to inaccurate and often unreliable facial feature tracking.

INCLUSION OF HUMAN BEHAVIOR INTERPRETATION IN DATA TAGGING AND RETRIEVAL LOOP

Only a few efforts have been reported so far on integrating the user's behavior in information tagging, seeking and retrieval process.

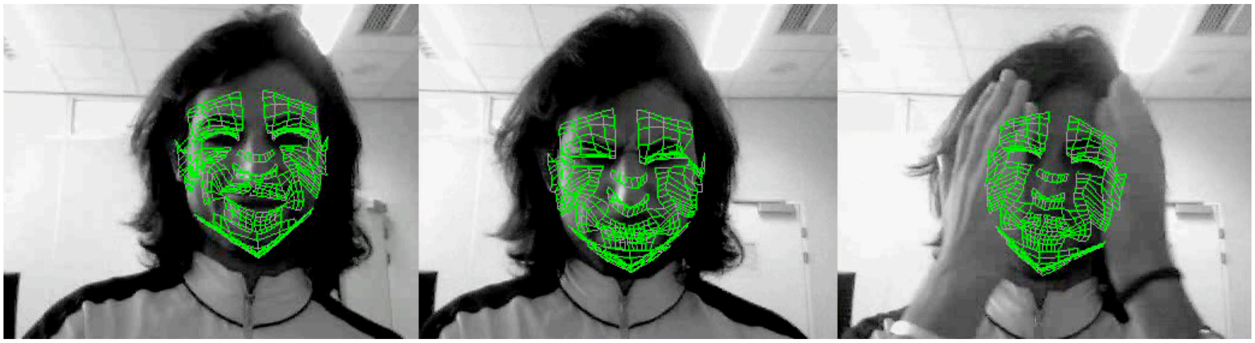

Figure 6: Results of the Piecewise Bezier Volume Deformation tracker (proposed in [14], used in [3]).

Petridis & Pantic [10] proposed a method for tagging video data in terms of hilarity of the watched video based on the user's laughter. The results suggest that, while laughter is a very good indicator of amusement, the kind of laughter (unvoiced laughter vs. voiced laughter) is correlated with the mirth of laughter and can be used to judge the actual hilarity of the stimulus data. For this study, an automated method for audiovisual analysis of laugher episodes exhibited while watching movie clips has been developed. The audio features based on spectral properties of the acoustic signal and the visual features based on facial feature tracking (see Figure 5) have been integrated using feature-level fusion, resulting in a multimodal approach to distinguishing voiced laughter from unvoiced laughter and speech. The classification accuracy of the system tested on spontaneous laughter episodes is 74%. The presented preliminary results provide evidence that unvoiced laughter can be interpreted as less gleeful than voiced laughter and consequently the detection of those two types of laughter can be used to label multimedia content as little funny or very funny respectively. The actual inclusion of an implicit tag, indicating the hilarity level of the watched video, into the retrieval process, has not been discussed by the authors.

Arapakis et al. [3] reported on a method that assesses the relevance of a video by analyzing affective aspects of the user's facial behavior. They used an existing method for automatic recognition of seven basic emotions (neutral, happiness, sadness, disgust, fear, anger, and surprise), which utilizes Bayesian network classifiers and facial features tracked by the Piecewise Bezier Volume Deformation tracker. This tracker employs an explicit 3D wire-frame model consisting of 16 surface patches embedded in Bezier volumes [14]. To learn affective aspects of the facial behavior typically observed when the user watches a relevant (or irrelevant) video, Arapakis et al. observed 24 users and the affective aspects of their facial behavior while watching various relevant and irrelevant videos. Based on the so obtained ground truth data, they trained a statistical binary classifier of the affective aspects of the observed facial behavior that predicts the relevance/ irrelevance of the currently watched video with an accuracy of 89%. Neither the definition of an implicit tag that could indicate the likelihood that the explicit tag associated with the target video is incorrect (i.e., that the watched video is irrelevant given the current query), nor how this information could be included to enhance the retrieval process, have been discussed by the authors.

Kierkels et al [6], presented a user-dependent approach to using affective information, extracted from the user's physiological reactions, as tags for multimedia content indexing and retrieval. They use a dimensional approach to affect recognition and classify the user's physiological reactions, including ECG and facial EMG signals, in terms of quantized values in valance-arousal (VA) space [5]. To train this classifier, they let 7 subjects watch 64 various video clips aimed at eliciting various affective states, asked the subjects to self assess their affective states in terms of a small number of quantized values in the AV space, and learned the mapping between the recorded bio signals and the self assessments. For multimedia tagging purposes, the user's bio signals were recorded and mapped into the VA space using the trained affect classifier. To achieve retrieval based on affective queries (e.g., retrieve "amusing videos"), a representation of the target queries in the VA space has been defined in the form of a Gaussian probability distribution, and the retrieval of the videos previously annotated with the resulting VA values has been implemented. Although the method is a promising first step towards inclusion of the user's affective behavior into the tagging and retrieval loop, the method achieved rather low precision, indicating that research on this topic and the corresponding technology is still in its pioneering stage.

CHALLENGES

Implicit, human-behavior-based tagging and retrieval systems could bring around a long-sought solution to flexible, general, non-tiresome, and statistically reliable multimedia tagging and retrieval. Yet, only few efforts have been made so far to include the observed user's reactions and behavior into the retrieval loop. Except of the fact that automatic analysis of human spontaneous reactions and behavior in front of the computer is far from being a trivial task, and the fact that a proper inclusion of implicit tags in the data tagging and retrieval loop is yet to be investigated, researchers in the IHCT field face a number of additional challenges.

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, the 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 (e.g., introvert persons are less likely to display their emotional reactions) and transient conditions like stress and fatigue that decrease the reactivity of the user. Although building culture-specific or user-specific methods could solve this, 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., funny, disgusting, horror, etc., or in terms of valance and arousal). Another important issue relates to the user's privacy and how to ensure that the observed user's behavior would be used only for data tagging and retrieval purposes, and not for building models of the user's behavioral patterns that could be misused for the purposes of advertising or surveillance.

In summary, defining a proper way of addressing all these issues, developing human behavior analyzers that can attain accurate and reliable results even when working with audiovisual sensors built in the commercial computers, and building safe and efficient human-behavior-based tagging and retrieval systems, open up exciting research avenues that remain to be explored.

ACKNOWLEDGMENTS

This work has been funded in part by the European Community's 7th Framework Programme [FP7/2007-2013] under the grant agreement no 231287 (SSPNet). The work of Maja Pantic is also funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

AUTHORS

Maja Pantic (m.pantic@imperial.ac.uk) is with Imperial College London, UK, Computing Department, where she is Reader in Multimodal HCI, and with University of Twente, the Netherlands, Department of Computer Science, where she is Professor of Affective and Behavioral Computing.

Alessandro Vinciarelli (alessandro.vinciarelli@idiap.ch) is a Senior Researcher at IDIAP Research Institute, Switzerland.

REFERENCES

[1] M. Ames & M. Naaman, "Why we tag: motivations for annotation in mobile and online media", *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pp. 971–980, 2007.

[2] N. Ambady & R. Rosenthal, "Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis", *Psychological Bulletin*, Vol. 111, No. 2, pp. 256-274, 1992.

[3] I. Arapakis, Y. Moshfeghi, H. Joho, R. Ren, D. Hannah & J.M. Jose, "Integrating facial expressions into user profiling for the improvement of a multimodal recommender system", *Proc. IEEE Int'l Conf. Multimedia & Expo*, pp. 1440-1443, 2009.

[4] R. Datta, D. Joshi, J. Li & J.Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age", *ACM Computing Surveys*, Vol. 40, No. 2, pp. 5:1-5:60, 2008.

[5] H. Gunes & M. Pantic, “Automatic dimensional and continuous emotion recognition”, *J. Synthetic Emotions*, 2009.

[6] J.J.M. Kierkels, M. Soleymani & T. Pun, “Queries and tags in affect-based multimedia retrieval”, *Proc. IEEE Int’l Conf. Multimedia & Expo*, pp. 1436-1439, 2009.

[7] K. Lerman & L. Jones, “Social browsing on Flicker”, *Proc. Intl. Conf. Weblogs and Social Media*, 2007.

[8] E. Murphy-Chutorian & M.M. Trivedi, “Head pose estimation in computer vision: A survey ”, *IEEE Trans. Pattern Analysis & Machine Intelligence*, Vol. 31, No.4, pp. 607-626, 2009.

[9] M. Pantic, “Machine analysis of facial behavior: Naturalistic and dynamic behavior”, *Phyl. Trans. Royal Society B*, 2009.

[10] S. Petridis & M. Pantic, “Is this joke really funny? Judging the mirth by audiovisual laughter analysis”, *Proc. IEEE Int’l Conf. Multimedia & Expo*, pp. 1444-1447, 2009.

[11] B. Reeves & C. Nass, *The media equation: How people treat computers, television, and new media like real people and places*, Cambridge University Press, 1998.

[12] B. Schuller, F. Eyben & G. Rigoll, “Static and Dynamic Modelling for the Recognition of Non-verbal Vocalisations in Conversational Speech”, *LNCS*, Vol. 5078, pp. 99–110, 2008.

[13] B. Schuller, R. Muller, F. Eyben, J. Gast, B. Hornler, M. Wollmer, G. Rigoll, A. Hothker & H. Konosu, “Being bored? Recognising natural interest by extensive audiovisual integration for real-life application”, *J. Image & Vision Computing*, Vol. 27, No. 12, 2009.

[14] H. Tao & T.S. Huang, “Connected vibrations – a model analysis approach to non-rigid motion tracking”, Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, pp. 735-740, 1998.

[15] M.F. Valstar, H. Gunes & M. Pantic, “How to Distinguish Posed from Spontaneous Smiles Using Geometric Features”, *Proc. ACM Int’l Conf. Multimodal Interfaces*, pp. 38-45, 2007.

[16] A. Vinciarelli, M. Pantic & H. Bourlard, “Social Signal Processing: Survey of an Emerging Domain”, *J. Image & Vision Computing*, Vol. 27, No. 12, 2009.

[17] M. Wollmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie, “Abandoning Emotion Classes - Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies”, *Proc. Interspeech*, pp. 597-600, 2008.

[18] Z. Zeng, M. Pantic, G.I. Roisman & T.S. Huang, “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions”, *IEEE Trans. Pattern Analysis & Machine Intelligence*, Vol. 31, No.1, pp. 39-58, 2009.