Vis Comput (2013)
DOI

ORIGINAL ARTICLE

A framework for implicit human centered image tagging inspired by attributed affect

Konstantinos C. Apostolakis · Petros Daras

Published online:

Abstract In this paper a framework for implicit human-centered tagging is presented. The proposed framework draws its inspiration from the psychologically established process of attribution. The latter strives to explain affect related changes observed during an individual's participation in an emotional episode, by bestowing the corresponding affect changing properties on a selected perceived stimulus. Our framework tries to reverse-engineer this attribution process. By monitoring the annotator's focus of attention through gaze tracking we identify the stimulus attributed as the cause for the observed change in core affect. The latter is analyzed from the user's facial expressions. Experimental results attained by a lightweight, cost efficient application based on the proposed framework show promising accuracy in both the assessment of topical relevance and direct annotation scenarios. These results are especially encouraging given the fact that the behavioral analyzers used to obtain user affective response and eye gaze lack the level of sophistication and high cost usually encountered in the related literature.

Keywords Implicit Human Centered Tagging · Affective Computing · Gaze Tracking ·

1 Introduction

As content databases are rapidly growing out of proportion, efficient means of implicit content annotation need to be defined in order to categorize and manage huge amount of data. Several methods reported in the scientific literature about implicit human-centered tagging (IHCT) indicate the use of user affective response (among and along others) as an ideal satisfaction metric [32]. User affective response can be obtained through traditional, explicit means (like textual annotation), using keywords such as "sad", "scary" and "disgusting", as well as implicitly, through the monitoring of user behaviour while they are engaged in an activity, such as a content search. The emotional factor is measured by analyzing a set of communicated signals, such as body language, facial expressions, gestures, voice pitch, heart rate and body temperature, all of which are subconsciously controlled body functions that are related to the user's current state of mind. Vinciarelli et al [40] documented the challenges posed by such an endeavour; mainly concerning the need to include the observed user reactions and behaviour (as well as the implicit tags themselves) to the data tagging and retrieval loop. In their work they argued that the development of behavioural analyzers, capable to attain both accurate and reliable results, even when the audiovisual sensors used to obtain behavioural information are mounted on today's commercial computers, is key to reaching that goal.

During content search within a large database of images, such as Google Images and Flickr, users may undergo several psychologically driven changes in state of mind, that are difficult to track or explain without proper identification of the emotion eliciting elements of the viewed content. A feeling of "disgust" for example, after looking at a specific image, can only be explained as the encounter with a specific stimulus that can only be credited as "disgusting" by the person doing the annotation. Russell [26] argued that attention behaviour is closely linked to such changes in psychological state of mind; the latter causing shifts of attention towards the objects attributed with the affect-changing properties in preparation for conscious and/or subconscious action. This process is the foundation of attributed affect, a concept derived from a dimensional psychological framework consisting of two primitives, namely core affect and the perception of affective quality.

In this paper we present a framework for implicit annotation of content inspired by the concept of attributed affect. We define an image tagging pipeline that allows users to directly annotate data based on a post-hoc explanation of core affect experience change via gaze monitoring. Our framework

K. C. Apostolakis · P. Daras ( )
Information Technologies Institute, Centre for Research & Technology Hellas, GR57001 Thessaloniki, Greece
e-mail: daras@iti.gr

therefore enables a novel and sensible approach to tagging data, reminiscent of the human point of view. Not only does a particular piece of data receive an appropriate personalized tag, but also, the reason behind this annotation preference is identified and can be utilized for further applications, such as automated tagging, recommendation and retrieval. We describe the translation of theory to practice by presenting a thorough documentation of a low-cost, accurate software application developed for implicitly tagging and assessing correctness of topical relevance in a large database of images. Through our tests, we record promising results that show our methodology can attain accurate tagging results that are comparable to the current state of the art, with the potential to be utilizable in many more applications.

The rest of this paper is organized as follows: Section 2 contains a complete overview of the proposed framework. This contains a summary of related work, an overview of the psychological framework from which we drew our inspiration and concludes with the contributions of this work. Section 3 describes the implementation details of an application developed for putting our framework and its applicability to cost-efficient behavioural analyzers to the test. More specifically, Section 3.1 concerns the facial feature point extraction procedure necessary for recognizing user core affect experience through facial expression analysis. Section 3.2 describes the integrated single image eye tracking system used to monitor user gaze behaviour, while Section 3.3 covers object recognition and extraction through the intuitive use of a popular foreground / background segmentation algorithm that has been modified to receive input by the eye gaze tracker. Section 4 gives an insight on the developed test application scenario and covers the experimental results that further reinforce our faith in the correctness of our framework. Finally, Section 5 concludes with a brief synopsis and an insight on possible improvements and extensions that can be made to the framework, which will serve as a basis for future work.

2 Framework overview

IHCT is a rather young research topic in which researchers try to envision ways for translating user interaction with multimedia data into sensible and effective annotation labels. These labels or tags are believed to improve the quality of organization and retrieval services [40]. As the process of implicitly tagging data stems from natural interaction with the content, which is neither “forced” or driven by user personal goals and motives, the resulting tags are expected to be more general and statistically robust, therefore more usable in contrast to explicit annotation methods, as is described in [32].

In the remainder of this Section we present the state of the art in current IHCT methodologies and applications and differentiate our work by thoroughly explaining the psychological framework that inspired it.

2.1 Summary of related work

In the related literature, implicit tagging has been used for direct annotation of data (such as images, video and music) with predefined sets of implicit tags (such as affective labels for describing emotion elicitation) [34] [38], assessment of explicit tag quality and correctness [1] [15] [33], user profiling by tracking personal preferences [11] and content summarization based on implicitly obtained feedback used mainly for re-ranking of results [41]. We limit our documentation of related work to research concerning the tagging of visual content such as images and videos, as they are more relevant to our image-oriented approach and experiments described in this paper, but we encourage readers to refer to the works of [2] [4] [28] concerning topical relevance of textual search results, as well as research on implicit characterization of musical scores [29], for a complete overview on emerging methodologies concerning IHCT.

Arapakis et al [1] utilized visual analysis of facial expressions as well as other physiological signals (such as galvanic skin response, body temperature, heat flux and accelerometers) for predicting relevance of video search results to a specific topic, achieving an accuracy performance of 66.5%.

In addition to affective tagging by utilizing visual and physiological signals, eye gaze has also been studied as a measure of relevance assessment and implicit tagging. Hajimirza et al [11] extract eye gaze features of users examining a set of images to assign a per-user level of interest to each image, which can be utilized for image tagging as well as for retrieval purposes. More specifically, the authors are able to extract a level of interest ranging from 0 (no interest) to 1 (fully interested) using a fuzzy logic based gaze inference system, reporting an accuracy of 53%. The potential of exploiting implicit gaze feedback data for the improvement of query-specific recommendations for movie clips is also explored in Vrochidis et al [41]. In their experiments using a content-based video search engine, they recorded past user gaze fixation and pupil dilation data using an eye tracking system in order to generate a set of features that describes each video being gazed at. Then, using an SVM binary classifier (relevant/non-relevant), these features are used to assess relevance of the video content to a query. The latter can then be used to recommend video results to similar queries posed by new users. The reported classification accuracy in their experiments using the full magnitude of the reported gaze features averages 85.7%, with a best reported result of 95.1%.

Jiao and Pantic [15] experimented with the use of facial expressions as a means to assess correctness of explicitly annotated tags to a set of images. Their work is based on the assumption that users are likely to display certain emotional cues (more notably, facial expressions) when confronted with correct or incorrectly annotated data. Geometric features are extracted from facial feature points during the experiment process and are fed forward to HMM binary classifiers. The results prove that facial expressions do convey information about user agreement or disagreement regarding the correctness of tags, reporting a per-participant prediction accuracy of 72.1%. Eventual research utilizing additional modalities, more prominently gaze behaviour highlighted the benefits of using more implicit feedback signals to cover those cases in which users do not convey their agreement or disagreement via facial expressions alone [33]. Soleymani et al [34] also experimented with a multimodal approach to annotate video data with affective labels utilizing gaze data and electroencephalogram (EEG). In their work, they defined ground truth for segments of emotional video clips classified explicitly via questionnaires under one of three classes for

both arousal (calm, medium, aroused) and valence components (pleasant, neutral, unpleasant). Then, the authors proceeded with automated generation of tags derived from these sets, via classification of bodily responses. Their best reported accuracy concerning the arousal dimension was 76.4%.

The use of facial expressions as a means of implicit annotation in comparison to explicit annotation techniques was also addressed by Tkalčič et al [38]. In this work the authors extract Gabor features from video frames depicting user facial expressions and use a k-nearest neighbors classifier to generate affective tags in the 3-dimensional emotional space of valence-arousal-dominance. In their experiments, the authors compare three methods of content annotation in terms of content-based recommender system performance. Although the scales were tipped in favour of the explicit annotation approach for recommendation, the authors felt their approach significantly improved content-based retrieval performance over generic metadata.

All of the afore-mentioned methodologies in the relative scientific literature agree upon the benefits of IHCT in terms of robustness and natural integration within the current data tagging and retrieval pipeline. Some of these works highlight the importance of fusing different implicit feedback signals to improve the quality of the results. Each modality contributes to the final annotation result as an independent measurement, which adds up to the total tally of relevance assessment, however no clear link is defined between the implicitly obtained feedbacks of the components. A comprehensive overview of the methods and their results is presented in Table 1.

Our framework presented in this paper is inspired by a more natural description of how human psychology generates responses to visual stimuli, by trying to explain the changes detected in core affect experience using gaze information. In order to elaborate on how we strive to achieve such naturally driven annotation scheme, we follow this report on the current state of the art with an overview of the psychological framework that inspired our work.

Study	Modality	Content	Best reported result
Arapakis et al [1]	Facial expressions, Physiological Signals	Video	66.5%
Hajimirza et al [11]	Eye gaze	Image	53.0%
Vrochidis et al [41]	Eye gaze	Video	95.1%
Jiao & Pantic [15]	Facial expressions	Image	72.1%
Soleymani et al [34]	Eye gaze, pupil	Video	76.1%

Table 1. Summary of related work and experimental results in the implicit video and image tagging scenarios.

2.2 Psychological Framework

Russell's definition of core affect [26] adopts the two-dimensional valence-arousal emotional space, and is defined as the experience of any given psychological state of mind, that can be represented as a tuple $C(v,a)$ inside a circular two dimensional space. The latter is comprised of a measure of valence (the amount of pleasure/displeasure experienced at any given time) and arousal (the level of activation in preparation for action). On the other hand, a stimulus' affective quality is defined as a property described in terms of valence and arousal, whose perception may or may not alter a person's current state (or experience) of core affect.

Throughout the course of an emotional episode, experience of core affect is prone to change. Any individual participant of such an event will subconsciously attempt to attribute such changes to their perceived causes. Each cause of change is linked to a single stimulus selected from a number of antecedent events preceding the aforementioned influence in core affect experience. The attributed antecedent becomes the subject of focus, henceforth referred to as the "Object", using Russell's terms. The "Object" is therefore firmly established as the cause of the whole experience. To clarify the attribution process with an example, the experience of boredom can be attributed by an individual to the attendance of a perceived boring event. Likewise the experience of fear, attributed to the perception of a terrifying threat. Here, "boredom" and "fear" define experiences of core affect while "boring" and "terrifying" are perceived affective qualities of stimuli attributed as the cause. It becomes clear that the affective quality measures the "Object's" ability to stimulate a specific emotional experience to a participant of an emotional episode. Furthermore, different participants may attribute similar experiences of core affect to different "Objects", or perceive the same "Object" with contradicting affective qualities.

The three features described above (change in core affect, the "Object" and attribution of the former to the latter) define the concept of attributed affect. Attributed affect is, in Russell's words, the most important of the concepts derived from the core affect framework primitives, as it defines emotional awareness. It is suggested that people will generally shift their attention to stimuli they believe to be somehow linked to their current state of mind. Attributed affect is the main route to the affective quality of the "Object", unveiling an individual's motivations for liking or displeasure felt towards certain objects and situations.

Our implicit content annotation framework draws its inspiration from the concepts of attribution and perception of stimuli affective qualities. We achieve this by monitoring user behavioural reactions to the viewing of content, manifested through gaze behaviour and the expression of emotion via facial expressions, in an attempt to discover its perceived affective quality. Through this reverse-engineering of the process of subconscious attribution, we obtain an appropriate sentimental tag describing content affective quality, as well as the "Object", perceived by the user as the cause for stimulating that state of emotion/response. Based on the assumption that the cognitive processes were triggered by the perception of the "Object's" affective quality, we can predict possible reactions of a particular user to similar content and proceed with direct annotation of content that has yet to be viewed by that individual (as an example, we can predict an arachnophobic's tendencies to tag every image depicting a spider as "scary" by monitoring that individual's reaction at the sight of a spider being depicted in a single image).

2.3 Contributions of this work

All of the aforementioned research works on IHCT report on the potential that this young research topic yields towards generating a care-free data annotation scheme that draws its strengths upon the robustness and generality of the actual tags themselves. However, none of the reported state of the art

methods concern themselves with a very simple yet important question: “Why does this piece of data make me feel the way I do?” It is our firm belief that establishing methods to obtain an answer to this question can open up new and exciting opportunities towards a number of open research fields closely partnered with IHCT, such as recommendation services and content-based image retrieval (CBIR).

The novelty of our proposal lies with the actual reverse engineering of the cognitive processes that drive our subconscious need to attribute our current state of mind to a perceived cause. Identifying the “Object” within an image provides a means to obtain the actual query beforehand, its affective quality used to generate an appropriate, completely personalized affective tag. Summarizing our thoughts to a list of potential advantages opened up by the concept of attributed affect, the proposed framework can be utilized to gain a certain number of advantages over the related literature surrounding the IHCT problem:

1. As users are more likely to exhibit similar reactions when perceiving the affective quality of a certain stimulus already identified by a past experience, automatic annotation of large portions of an image database is possible by allowing users to browse and look at only a limited number of images. We already touched on this subject with our previous example of an arachnophobic's unpleasant encounter with an image depicting a spider. This antecedent can lead to all images of spiders within a database to be annotated as “unpleasant” for this particular user.
2. Retrieval and recommendations can be made readily available through the annotation of the specified stimuli. Returning to our example, relevant results during a content search would consider the “unpleasantness” of spiders and rank all related images with lower scores in an attempt to avoid a negative interaction of the system with the particular user.
3. Annotation is based on user personal experience, which further addresses culture-dependent annotation problems. What may register as “funny” in one culture could be considered offensive to another. Our framework addresses these issues, as annotations are made personal by each individual, the cause for each tag preference stored within the concept of the “Object”.
4. Our description of the proposed framework is general and therefore open for experimentation with a number of methodologies and hardware components already proposed in the related literature. This ensures an unprecedented scalability of our approach in terms of cost, accuracy and real-time performance of the employed IHCT application.

3 From theory to practice

The premise of how the proposed framework includes the implicit affective tags into the data tagging and retrieval loop closely resembles the description of attribution in the psychological background presented in Section 2. An application was then built to realize the framework as an actual piece of software addressing IHCT requirements with an emphasis on modern and portable hardware solutions. Our development goals were twofold: design a novel workflow that would simulate the reverse-engineering of the attribution process, while relying on cost-efficient, yet accurate and reliable behavioural analyzers utilizable by the average user. In the remainder of this Section we elaborate on the implementation specifics of each module contributing to the final framework application.

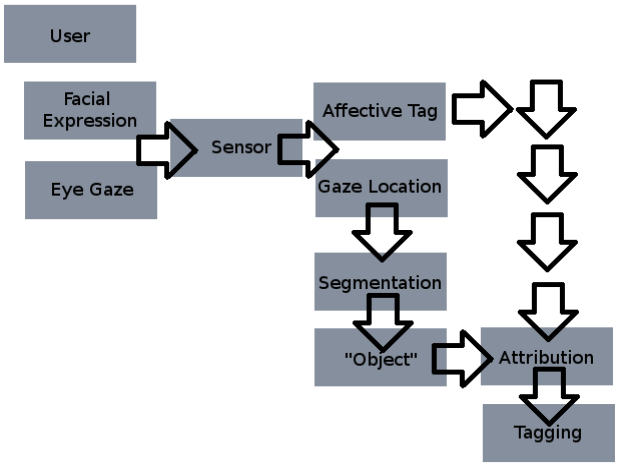

Fig. 1 Block diagram of the proposed framework

3.1 Overall framework architecture

From the previous Sections it has been made clear that the framework relies on user focus of attention in combination with affective response to generate the implicit emotional tag and infer the "Object" to which it is attached to. Application-wise, this structure requires implicit user input obtained from a sensor monitoring user activity during image browsing. Similar to the works of [2] [15] [33] and [38] we chose facial expression analysis as our main means to detect core affect experience during the interaction with the content. Also, we employ the use of a gaze tracking system to determine user focus of attention. These input signals (user facial expression and eye gaze) are processed to generate an appropriate affective tag as well as a gaze location on the screen. The latter is fed forward to a quasi-novel approach to identifying the “Object” from the acquired gaze data via image segmentation. Once the "Object" is extracted, it is attributed with the affective quality corresponding to the affective response recognized earlier. The eventual affective tag therefore refers to the specified "Object", rather than the general content of the image as a whole. A block diagram of the framework approach and different modules comprising the developed application is shown in Figure 1.

3.2 Affect recognition module

Affect recognition is an enormous research field in its own respect, with a plethora of methods and applications having been developed towards accurate and reliable analysis of user affective state. Thoroughly reporting upon the state of the art in this research field is beyond the scope of this paper, we instead encourage interested readers to refer to [44] for a summary of recent developments. The following paragraph clarifies implementation choices and details behind the affect recognition module developed for obtaining user affective response within our application framework.

Common in most facial expression analysis methodologies for

decoding visual information into sensible data is the work of Ekman & Friesen [10] concerning the Facial Action Coding System (FACS). Basically, FACS deconstructs every anatomically possible facial expression into a set of so called Action Units (AUs) that describe the movement of individual facial muscles during the display of a specific facial expression. Different combinations of AUs produce different facial expressions. Detection of muscle group displacement is possible by identifying a number of facial feature points and tracking their location during the display of an expression. A number of different approaches have been proposed in the scientific literature for locating such landmarks, with Viola-Jones cascade classifier detectors [39], Active Shape and Appearance models [7] [8] and model tracking approaches [37] being among the fastest and most popular solutions. In fact, most of these algorithms have been developed to the point of maturity, which has allowed their integration into a number of robust software libraries such as OpenCV[1] and asmlibrary[2].

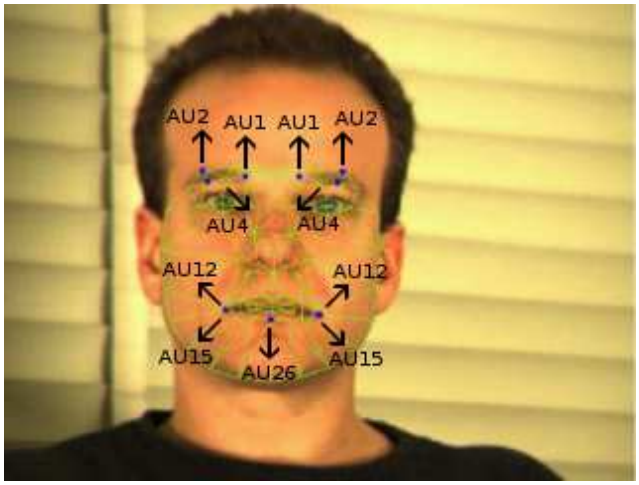

Fig. 2 Facial landmarks tracked for AU activation identification, showcased on an image example of the IR Marks face database [45].

3.2.1 Method

For our framework application, we handled the identification and tracking of facial feature points by first detecting the user's face and then localizing a set of key facial features. To this end, we used a Viola-Jones classifier cascade [39] for face detection, and an Active Shape Model (ASM) fitting algorithm for landmark tracking. An ASM is a statistical model developed by Cootes and Taylor [7], describing the shape of an object. It is capable of deforming, in order to fit to a new instance of the object presented in an image (or, as is the case here, a camera frame). The ASM algorithm tries to match the trained model (basically, a set of points representing the shape of an object) to a new image by iterating over the following steps:

a) Re-locate the model points to better found locations close to the original fit of the model to the image.
b) Update the model's parameters to better match the relocated points.

Through the ASM fitting procedure, the location of each of the feature points is extracted via its corresponding landmark in the model's shape. The method is highlighted in Figure 2. The set of identified AUs includes the inner brow raiser (AU1), outer brow raiser (AU2), brow lower (AU4), lip corner puller (AU12), lip corner depressor (AU15) and jaw drop (AU26).

3.2.2 Mapping Action Units to core affect space

The extracted AUs describe the affective response during the currently experienced emotional episode (reminiscent of a familiar emotion class, like for example "Fear") and are therefore linked to an affective term defined in core affect space. This latter term, is the actual tag used to annotate the data. Martinez Bedard's study [20] makes a clear distinction between the quadrants of core affect circular space, stating that certain clusters of properties, including facial cues provided by corrugator and zygomatic muscles, respectively controlling the brow and mouth area AUs previously extracted, are associated with affective terms falling within a single quadrant. Further works of Yik [42] and Yik et al [43] provide a useful insight by splitting the core affect circular space into 12 segments, each one associated with a set of sentiment tags falling under the same category.

To the best of our knowledge, a direct and generally approved mapping function of AUs to core affect tuples has yet to be adopted by the scientific community. In fact, psychological literature is rich and diverse with opinions and arguments about the nature of emotion, and its relation to facial expression display is an open subject of discussion. In this work we try to follow some of the more generally accepted hypotheses, both within the psychologist and affective computing expert literature, with respect to the fact that these hypotheses may be challenged by researchers supporting a different mapping concept.

Our initial mapping efforts stem from the works of Smith and Scott [31] as well as Simon et al [30]. The latter provides mean ratings of valence and arousal levels corresponding to facial expressions posed by both male and female actors. Intensity of the measurements ranges from -4 to +4 for both core affect dimensions. Expressions are coded by AU combinations based on Ekman's prototypical emotion representation and actual observations of AU activation whenever an expression is displayed by an actor. This valence-arousal mapping to AU activation during certain facial expressions is shown in Table 2. Following a similar simplified analysis scheme as in [35] [15] and [33], our mapping scheme is composed by monitoring landmarks tracking AU displacement in the mouth and brow areas of the face. Our conclusions are further supported by the published works of Huang [12], concerning emotion recognition via ASM fitting, as well as the work of Lim & Aylett [18], where valence and arousal components are mapped to mouth and eyebrow movements of a cartoon face.

3.2.3 Implementation details

We modeled the shape of the human face displaying a varying set of emotional facial expressions using the 68-point facial landmarks defined in [9] and shown in Figure 3. The model was trained on 161 manually annotated frontal face images obtained from various sources [22] [21] [14] [19]. Our AU identification scheme closely resembles that of [35], and is

[1] http://opencv.org/
[2] https://code.google.com/p/asmlibrary/

Facial Expression	Corresponding Action Units	Mean valence estimate	Mean arousal estimate
Happiness	**6**+**12**	+2.990	+2.140
Anger	**4**+7+23	-1.685	+1.240
Fear	**1**+**4**+5+25	-2.215	+1.475
Surprise	**1**+**2**+**26**	-0.010	+1.515
Sadness	**1**+**4**+**15**	-2.190	-0.605
Neutral	-	+0.025	-1.205

Table 2. Estimated Valence-Arousal mapping to AU activation during posed display of certain facial expressions reported in [30]. In bold are the AUs actually identified by our affect recognition module.

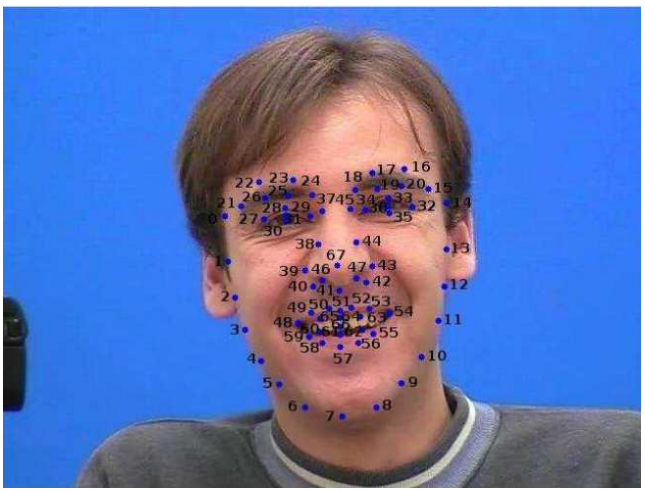

Fig. 3 The 68-point landmark annotation scheme used for training the Active Shape Model

achieved by measuring the distance of each selected feature point to the fictional line connecting the inner eye corners, henceforth referred to as the eye line. This distance is also measured for each feature point during the display of a neutral, unemotional expression, obtained in an offline step using a single frame depicting the user in an unemotional state, providing a means to identify if and when each muscle group has been activated. The furthest away the feature point moves from the eye line in relation to its original location, the greater the magnitude of the AU effect on the resulting facial expression.

Most of the AUs present in Table 2 (with the exception of 26) come in pairs, and therefore, both left and right counterparts have to be considered when calculating AU intensity. All valence-arousal pair values are normalized inside the [-1, 1] numerical space. Through an analysis of the results presented in the aforementioned literature, we estimate the valence component through AUs 4, 12 and 15, according to the following equation:

$$v = AU12 - \left(\frac{AU15 + AU4}{2} \right) \qquad (1)$$

Meanwhile, the arousal component gradually rises from its low starting neutral level (which is normalized at -0.30125) as the intensity estimates for AUs 1, 2, 4, 12 and 26 steadily increase, as is described in the following equation:

$$a = 1.30125 \cdot \frac{a^{+}_{sum}}{5} - (AU15 + 0.30125) \qquad (2)$$

	excited	elated	satisfied	relaxed	serene	still	sluggish	sad	unhappy	irritated	enraged	awed
excited	0.03	0.43	0.38	0.08	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.03
elated	0.01	0.21	0.62	0.09	0.03	0.03	0.01	0.00	0.00	0.00	0.00	0.00
satisfied	0.01	0.17	0.31	0.18	0.11	0.16	0.04	0.00	0.00	0.01	0.01	0.00
relaxed	0.02	0.13	0.22	0.09	0.08	0.38	0.05	0.01	0.01	0.01	0.00	0.00
serene	0.01	0.06	0.13	0.07	0.08	0.44	0.15	0.04	0.02	0.00	0.00	0.00
still	0.00	0.02	0.09	0.06	0.06	0.44	0.07	0.06	0.12	0.01	0.00	0.07
sluggish	0.04	0.04	0.12	0.06	0.06	0.41	0.08	0.05	0.13	0.01	0.00	0.00
sad	0.00	0.05	0.21	0.11	0.06	0.22	0.09	0.10	0.15	0.01	0.00	0.00
unhappy	0.00	0.11	0.26	0.08	0.05	0.16	0.06	0.06	0.18	0.04	0.00	0.00
irritated	0.02	0.06	0.07	0.06	0.06	0.14	0.03	0.02	0.28	0.22	0.02	0.02
enraged	0.04	0.06	0.05	0.06	0.06	0.12	0.02	0.03	0.16	0.29	0.07	0.05
awed	0.05	0.01	0.09	0.02	0.02	0.05	0.03	0.03	0.07	0.01	0.00	0.61

Table 3. Confusion matrix for the affect recognition module. Note how most of the affective terms are accurately placed within the neighboring segments of the core affect circular space (green band).

where:

$$a^{+}_{sum} = AU1 + AU2 + AU4 + AU12 + AU26 \qquad (3)$$

Each valence-arousal pair is then placed within one of the 12 core affect segments mentioned in [43], each one associated with an eventual affective tag. Out of these tags, five reveal pleasantness in perception with varying levels of arousal (*serene*, *relaxed*, *satisfied*, *elated*, and *excited*). Another five labels correspond to negative feedback, also dependent on the current level of the arousal value (*sluggish*, *sad*, *unhappy*, *irritated* and *enraged*). The remaining two labels conveying no strong valence-related information neither in the positive or negative direction, accounted for highly activated (*awed*) and neutral (*still*) feedback. These groups form the 4 kinds of affective feedback classes used for experimental validation of our framework.

3.2.4 Validation

We put the accuracy of our affect recognition module to the test by asking users to pose facial expressions corresponding to the 12 core affect segment descriptions of [43], recording the affective tag output of the module for each posed expression. The affect recognition module generates an affective term in real time. As can be seen in the confusion matrix shown in Table 3, neighboring affective terms in the core affect circle are confused with one another, maintaining consistency within the core affect categories defined in the previous Section (positive, negative, aroused and neutral feedback), which validates the effectiveness of the approach.

3.3 Gaze tracking module

Facing the challenge of developing highly accurate and cost-efficient analyzers utilizable by the average user, we decided to conform our gaze tracking system with the sensory input already deemed sufficient for our affect recognition module. Commercially available, high precision gaze-tracking systems such as Tobii[3], SMI[4] and EyeTech[5], were therefore not considered as candidates for integration. Instead, an image processing-based approach that makes use of a single camera was followed for performing gaze-tracking in this work.

3.3.1 Method

Most traditional image processing-based gaze-tracking systems rely on the detection of distinct facial features, such as the eye corners or the pupils. Localization of such features has already been addressed in our affect recognition module described in the previous paragraph. Each of these aforementioned approaches estimates a consequent gaze vector by either incorporating geometrical models of the human eye [13][24][6][17] or by using mapping functions for relating gaze parameters to screen coordinates [45]. The latter methods usually require user-specific data to be collected via an off-line calibration procedure.

Our gaze tracking module utilizes the ASM to localize useful features around the user's eye areas, such as the eye corners. These eye areas are processed in order to extract the pupil centre. We adopt the method of [45] to extract a gaze location on the screen. More specifically, we employ linear 2D mapping of eye corner-to-pupil centre vectors to a corresponding pair of screen coordinates, a process that requires an off-line calibration procedure to be performed per subject, prior to the use of the tracker.

3.3.2 Tracker calibration

This procedure constitutes the association of eye corner $E(\chi,\psi)$ to pupil centre $P(\chi,\psi)$ vectors for each eye to a set of known screen coordinates (called calibration points) being successively displayed on the screen. We used a total of eight calibration points placed diametrically on the screen boundaries, as is shown in Figure 4. Once calibration of the gaze tracker is complete, estimating the gaze point on the screen follows the practice of [45] through a linear mapping of coordinates between the camera frame image plane and the screen.

More specifically, for each eye an eye corner to pupil centre vector $U_i(\chi_i,\psi_i),\ i \in [1,8]$, is stored for each of the eight calibration points that correspond to the eight aforementioned screen boundary locations $\Delta_i(x_i,y_i)$. In the latter notation, the coordinates (x,y) refer to the computer screen image plane. In [45], only two calibration points are required for the definition of the mapping function, namely the values of variables $(\chi_{right}, \chi_{left}, \psi_{top}, \psi_{bottom})$, which correspond to

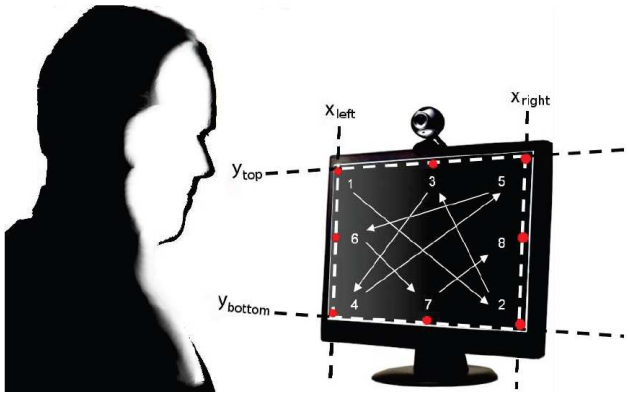

Fig. 4 Gaze-tracking framework setup and calibration procedure.

the known values $(x_{right}, x_{left}, y_{top}, y_{bottom})$ defined in Figure 4. To reinforce robustness of our gaze tracking module, we exploited the information from the eight aforementioned calibration points to estimate these values, as is described in the following equations:

$$\chi_{left} = \frac{\chi_1 + \chi_4 + \chi_6}{3} \quad (4)$$

$$\chi_{right} = \frac{\chi_2 + \chi_5 + \chi_8}{3} \quad (5)$$

$$\psi_{top} = \frac{\psi_1 + \psi_3 + \psi_5}{3} \quad (6)$$

$$\psi_{bottom} = \frac{\psi_2 + \psi_4 + \psi_7}{3} \quad (7)$$

3.3.3 Implementation details

For the development of our gaze tracking module, we chose a combination of fast and reliable methods reported in the previously mentioned related literature, looking to take advantage of the ASM landmark features already employed for facial expression analysis in the affect recognition module. As previously noted, we utilize these landmarks to isolate regions of interest in the camera frame that contain the eye areas. These regions are used to localize the pupil centre, by applying an adaptive thresholding technique, similarly to the well-known Otsu histogram shape-based image thresholding algorithm [23]. This technique iteratively applies increasing threshold values to the gray scale converted images of the eye areas, producing binary images in an attempt to segment the darker pupils from the lighter-toned background (e.g. sclera). The pupil center is then located by calculating the medium point $P(\chi,\psi)$ of the collection of points making up the extracted pupil object's contour. In this particular notation, the coordinates (χ,ψ) refer to the camera frame image plane. The overall pupil center tracking process is illustrated in Figure 5.

With the pupil centre point known, and eye corner feature points localized via the ASM, we proceed with the estimation of the gaze location on the screen. Utilizing the information obtained in the calibration step, the eventual linear mapping of an arbitrary eye corner to pupil centre vector $U(\chi,\psi)$ to a corresponding gaze point location on the screen $\Delta(x,y)$ conforms to the following equations:

[3] http://www.tobii.com/

[4] http://www.smivision.com/en.html

[5] https://www.eyetechds.com/

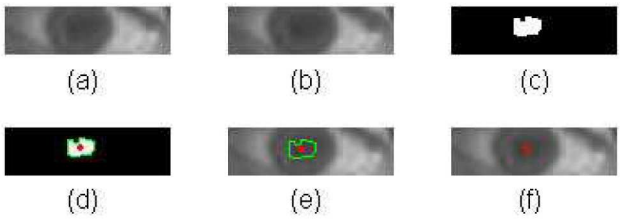

Fig. 5 Iris center detection process via automatic adaptive thresholding. a) The original image of the eye is converted to greyscale. b) The grayscale image is eroded. c) Increasing thresholds are applied in an iterative fashion to segment the darker pupil area from the background. d) The extracted object's contour is found. e) The median of the extracted object is calculated. f) The latter point is returned as the iris center.

$$x = x_{left} + \frac{\chi - \chi_{left}}{\chi_{right} - \chi_{left}} \cdot (x_{right} - x_{left}) \quad (8)$$

$$y = y_{top} + \frac{\psi - \psi_{top}}{\psi_{bottom} - \psi_{top}} \cdot (y_{bottom} - y_{top}) \quad (9)$$

Acquiring a gaze point $\Delta(x,y)$ for each of the subject's two eyes, the final gaze point position on the screen is computed by calculating the average of the coordinates of these two points.

3.3.4 Validation

We are not aware of any universally acclaimed method for measuring gaze tracker accuracy. Therefore, in order to validate our gaze tracker performance we defined a significantly challenging experiment that can be easily reproduced and takes both spatial accuracy and temporal coherence of the tracker into account. More specifically we asked users to follow with their eye gaze a red circle traversing a circular trajectory. The distance of the user's head was maintained at 65cm from the screen, while the radius of the red circle was set to 0.7cm. The circular trajectory radius was set to 13.5cm, and was fully traversed in 30sec. The tracker's accuracy was defined as the mean gaze angle deviation corresponding to the distance of the estimated gaze point to the red circle centre. In this way the mean angular error was calculated at 0.83 degrees. We performed a comparative study with the work of [45], which we used as a benchmark, as it is the closest related single camera gaze tracking scheme to our own. As this benchmark method reports a mean angular error of 1.4 degrees, we assert our tracker's efficiency.

3.4 "Object" extraction module

One of the major contributions of this work is the identification of the emotion eliciting "Object" contained within the image, deemed capable to inspire certain feelings as it's being gazed at by a user-annotator. Since we are interested in specific object extraction, segmentation algorithms enter the equation. The proposed framework's approach to image segmentation provides a simple and effective segmentation scheme for rapid foreground object extraction using eye gaze as input. We are aware of a single related effort, namely Sadeghi et al [27]. This work presents an interactive image segmentation system interface that allows users to designate specific foreground and background seeds by fixating their eye gaze on certain image locations. Their experimental results confirm that eye gaze information can effectively and tirelessly substitute explicit mouse interaction.

In order to keep the affective tagging and retrieval procedure as less intrusive as possible, several of the aforementioned interface options would have to be simplified. The explicit designation of foreground and background seeds might risk taking up more time than the annotation procedure itself, shifting the user's focus towards achieving a plausible segmentation of the object, instead of actually labeling the latter with an appropriate affective tag. Furthermore, such an elaborate procedure suffers the risk of tampering with the actual affective response (users may for example get frustrated during the segmentation process). Ideally, we believe the user should just look at the object depicted in the image, and the underlying segmentation procedure should take over the rest.

3.4.1 Method

As discussed in the previous paragraph, gaze point estimation serves as a sort of a visual mouse pointer being driven on screen by user eye gaze. When this intuitive mouse pointer is found to intersect one of the images displayed, a Region of Interest (ROI) can be automatically generated around that pointer, encapsulating a portion of the image where the user is assumed to focus his/her attention on. The "Object" is then extracted by segmenting foreground-background pixel data using the GrabCut algorithm.

GrabCut [25] is an interactive foreground object extraction algorithm that demonstrates exceptional extraction quality on complex background environment depictions, while requiring minimal user effort on its behalf. A simple implementation of the algorithm is documented in [36]. In its simplest form, the algorithm requires of the user to simply specify an area around the foreground object of interest. In the algorithm's initial run, image pixels located outside this rectangular area are marked as certain background pixels, forming a background class. Pixels located inside this area are similarly assigned as certain foreground pixels. The algorithm then re-assigns foreground pixels according to Gaussian Mixture Models (GMMs) constructed in the previous step in an iterative fashion, until the foreground/background classification converges. Depending on the level of segmentation quality required, recent modified versions of the classic GrabCut segmentation scheme [5] have improved segmentation accuracy.

3.4.2 Validation

In order to validate our "Object" extraction module, we used a Bag of Features pipeline [16] to classify each segmented "Object" image to its corresponding class. The actual image database we used for the experiments was collected in-house, and is comprised of the most frequently-appearing distinct image categories returned by Google Images on a search with the keyword "Paris". This database, which will hence forward be referred to as the *Paris* database, consists of a total of 1125 images, depicting several of the French capital's most famous landmarks, namely the Eiffel Tower, the Notre-Dame church,

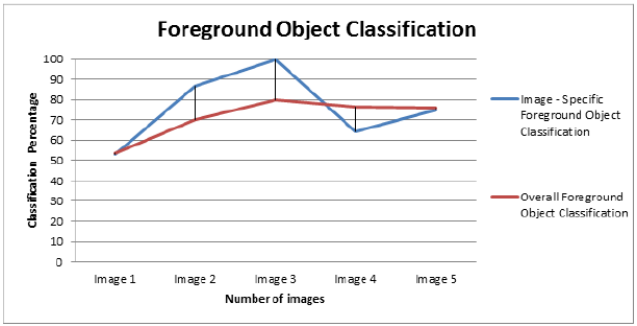

Fig. 6 Image-specific and overall foreground image classification rates for multi-image case experiment. Image 1 corresponded to the 'Eiffel Tower' class, Image 2 to the 'Notre Dame' class, Image 3 to the 'Celebrity' class, Image 4 to the 'Arc de Triomphe ' class and finally Image 5 corresponded to the 'Louvre Museum' class.

the Louvre and the Arc de Triomphe, as well as images depicting photographs of persons apparently sharing the same name. Users were shown 5 images, each of a separate category, and were told to look the images.

An "Object" image is basically an alpha matte image depicting only the foreground segmentation result of the GrabCut algorithm. The image is obtained after a user has visited the image with his/her gaze. SURF descriptors [3] were used for image feature extraction. A Radial Basis Function Support Vector Machine (SVM), trained using the complete “Paris” database, was used for the eventual classification of foreground images. Approximately 95% of the foreground images produced were actually usable for classification. As can be seen from the experimental results presented in Figure 6, the overall classification accuracy of the foreground objects approximately reaches 76%.

4 Test Scenario

The test scenario devised to put our framework application to the test touches on two of the categories of IHCT applications mentioned in [32]. First, we address the direct annotation of images with automatically generated affective tags. Secondly, we examine the assessment of topical relevance of the displayed content with the concept it’s being linked to. Through this scenario we also attempted to simulate a natural process of content search. Users were told to react towards the results returned by a hypothetical search engine in response to the query keyword "Paris". Content depicted in the images being gazed at would instantly be annotated with one of the twelve representative affective labels mentioned in [43].

The experiment was divided into two discrete steps, the first one concerning the display of a single, random image of the database. In the second step, each user was shown five randomly selected images of the *Paris* database contained within a single screen, each one corresponding to a distinct category of the ones defined within the *Paris* database. A screenshot is shown in Figure 7. When subjects concluded an experimental run, they were asked to explicitly assess appropriate tags for the images they looked at, using the aforementioned twelve affective terms. This last step was included to generate ground truth for assessment of the automated IHCT application results.

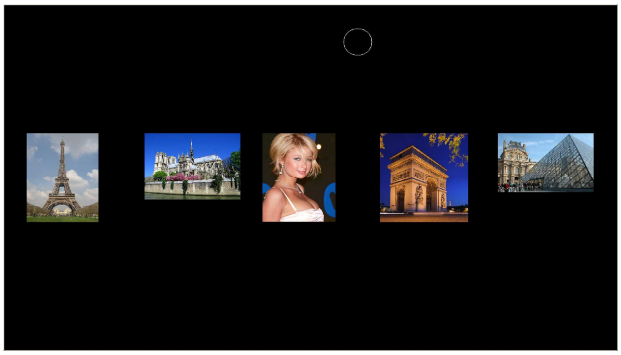

Fig. 7 Screenshot of the application in the multiple image display case. The white circle indicates gaze location.

4.1 Implementation details

The application interface implementing our proposed framework was developed in the C++ programming language, utilizing the computer vision resources mentioned earlier in Section 3.2. The application test environment was installed on a 3.30 GHz Intel Core i5-2500K desktop computer with a 24 inch Samsung SyncMaster 2494 display and a Unibrain Fire-I 1.2 fire wire webcam mounted on top. Display resolution of the display was set to 1280x768 pixels, while images, all of which originated from the *Paris* database, on display were forced to a maximum of 200 pixels per dimension, maintaining aspect ratio. Webcam resolution was set to 640x480 pixels with an output stream of 15 Frames per Second (FPS). A total of 15 healthy individuals, 13 male and 2 female aged between their early 20s to mid 30s, all expert computer users with average to non-existent experience of use of remote eye tracking systems volunteered as test subjects. All subjects were shown a tutorial video of approximately 3 minutes length, and received thorough instructions on using the application interface and the aim of the test scenario. The users were first instructed by the application to display a neutral facial expression as explained in Section 3.2.3. Then, they were able to proceed with calibration of the eye tracker, as mentioned in Section 3.3.2. Both steps were similar in both cases. After neutral face pose was captured, an indicator showcasing current core affect as a point inside core affect circular space was made visible to the users. After calibration of the eye tracker, gaze point location on the screen appeared as a hollow circle with a radius of 30 pixels.

5 Results and discussion

5.1 Results

We accumulated our experimental results by comparing the automated IHCT framework application classification results to the explicitly obtained ground truth. We categorized the affective terms of [43] into the 4 classes representing positive, negative, activated and neutral feedback, as described in Section3.2.3. The eventual results concern our two-fold goal in addressing the IHCT application topics. The single image case results shown in Figure 8, concern the assessment of topical relevance of the depicted content to the keyword "Paris". Our framework is shown to achieve an approximate 80% correct classification rate into one of the four mentioned categories. More specifically it is shown that out of the 15

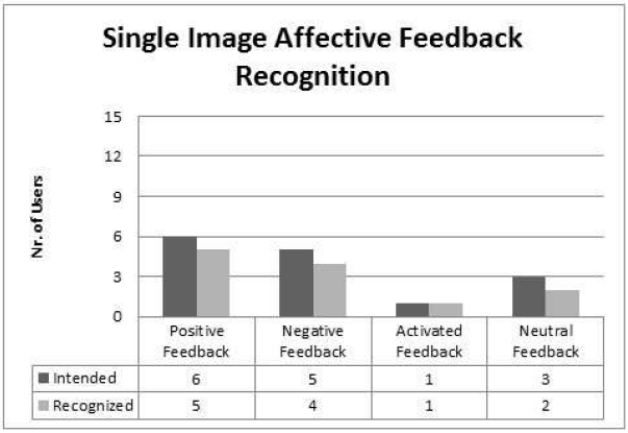

Fig. 8 Affective feedback classification results for single-image case experiment.

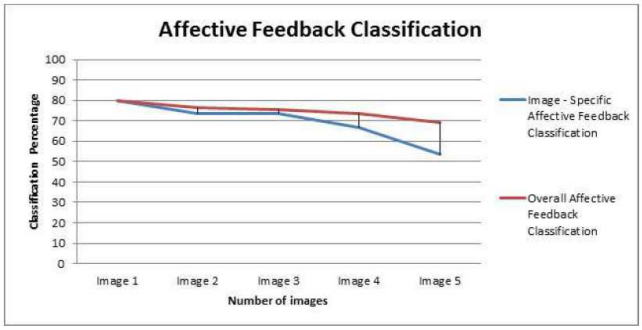

Fig. 9 Image-specific and overall affective feedback classification rates for multi-image case experiment.

explicitly designated relevance feedback classifications, our application managed to correctly classify 12 nonverbal reactions to the intended feedback class.

Figure 9 showcases the affective feedback classification results into the 4 affective feedback categories for the multi-image display case. As can be seen, the mean classification rate with respect to the ground truth suffers an approximate 10% loss to the single image case. Overall our IHCT pipeline performance achieved an approximate 70% correct affective feedback assessment of tags in the multi-image scenario case.

5.2 Discussion

We find the experimental results extremely encouraging, as they are comparable to the results reported in the related literature as summarized in Section 2.1, but were however attained in a much more efficient test environment with respect to cost and overall complexity.

To showcase the strength of our framework, we combine the results of these experiments with the "Object" extraction module classification rates reported at 76% back in Section 3.4.2. A closer look on Figure 6 indicates that the distinct 'Celebrity' category foreground images were 100% correctly classified, whereas all the other categories belonging to buildings containing similar feature patches (such as the blue sky for example) were sometimes confused. An interesting fact for discussion stems from the ground truth, where all users associated the 'Paris' keyword with most of the building categories while dismissing the 'Celebrity' category as irrelevant. As can be seen by closely examining Figure 9, approximately 73% automated feedback classifications for the 'Celebrity' category would correctly associate the entire category of images with negative tags as an indication of topical relevance to the keyword. Therefore, an endeavor to recommend future results to the same query could avoid the entire class of 'Celebrity' images altogether. This is especially interesting, as most users reported they associated the keyword with the French capital, rather than a person.

In Figure 9, we note a significant drop of classification rates as subjects rated images depicting Arc de Triomphe and Louvre Objects as mostly indifferent to the query in the ground truth. This, along with the fact that many of the images undergone segmentation proved to be unsuitable for "Object" classification, as users did not spend much time gazing at theses images, may attribute to the low classification rates for these specific images which did influence the final mean classification results. Furthermore, the majority of the test participants reported they rarely display clear visible facial signs of approval (such as a smile or a grin) during the viewing of keyword-related content, lest it be considered either funny or cute. Most subjects went on reporting however on the certainty of displaying clear signs of disapproval whenever encountering displeasing, irrelevant, or possibly offensive depictions of unpleasant content. Emotional displays corresponding to high arousal states (e.g. surprise) were also deemed to be more likely to occur than not, considering the intensity of the surprising, exciting or shocking quality associated with the image content. As such, other means of extracting user affective response could complement, or substitute facial expression analysis and possibly generate more promising results. This is a subject for further research.

Also, of particular interest were the results emerging from this study showing images depicting multiple “Objects” led to classification results depending on the user’s attention behaviour. In one such example, a user who actually looked at the vague shape of a distant Eiffel Tower in the background of an image clearly taken of the Arc de Triomphe, actually associated his affective response with all images containing the Eiffel Tower. Such results reinforce our claim on the potential our framework has for application in content retrieval and recommendation scenarios.

6 Conclusions

In this paper, we introduced a framework for implicit human-centered tagging inspired by the concept of attributed affect, focusing more on the cognitive process that leads to the association of an affective label with a certain “Object” that is present in the scene. We have argued that reverse-engineering this process into a data tagging pipeline that utilizes affect recognition and gaze tracking modules, can attain accurate direct tagging and topical relevance results. Also, the extra information on the “Object” itself can open up new opportunities in the partnered fields of recommendation and content based retrieval. Our theory has been put to the test through an integrated IHCT application built around low-cost yet accurate behavioural analyzers, showcasing the potential of our proposed annotation scheme.

Further research work related to this framework includes a comparative study between approaches reported in the scientific IHCT literature as well as putting our hypotheses on the benefits gained by using our framework in

recommendation and CBIR use case scenarios. In this respect, we have every reason to believe our framework can serve as a basis for combining some of the more sophisticated affect recognition and gaze tracking applications proposed in the related literature (such as the acquisition of physiological signals) and speculate on the potential contribution to an increased accuracy in direct annotation scenarios.

Acknowledgment

The research leading to this work has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2011-7-287723 (REVERIE project).

References

[1] Arapakis, I., Konstas, I., Jose, J. M.: Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In: Proceedings of the 17th ACM International conference on Multimedia (pp. 461-470), ACM, 2009.

[2] Arapakis, I., Athanasakos, K., Jose, J. M.: A comparison of general vs personalized affective models for the prediction of topical relevance. In: Proceedings of the 33rd International ACM SIGIR conference on Research and development in information retrieval (pp. 371-378), ACM, 2010.

[3] Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Computer Vision-ECCV, 2006 (pp. 404-417), Springer Berlin Heidelberg, 2006).

[4] Buscher, G., van Elst, L., Dengel, A.: Segment-level display time as implicit feedback: a comparison to eye tracking. In: Proceedings of the 32nd International ACM SIGIR conference on Research and development in information retrieval (pp. 67-74), ACM, 2009.

[5] Chen, D., Chen, B., Mamic, G., Fookes, C., Sridharan, S.: Improved grabcut segmentations via GMM optimisation. In: Computing: Techniques and Applications, 2008, DICTA '08, Digital Image (pp. 39-45), IEEE, 2008.

[6] Chen, J., Tong, Y., Gray, W., Ji, Q.: A robust 3D eye gaze tracking system using noise reduction. In: Proceedings of the 2008 symposium on Eye tracking research & applications (pp. 189-196), ACM, 2008.

[7] Cootes, T. F., Taylor, C. J., Cooper, D. H., Graham, J.: Active shape models-their training and application. In: Computer vision and image understanding, 61(1), 38-59, 1995.

[8] Cootes, T. F., Taylor, C. J.: Active appearance models. In: Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23(6), 681-685, 2001.

[9] Cootes, T. F.: Talking Face Video.

[10] Ekman, P., Friesen, W. V.: Facial action coding system: A technique for the measurement of facial movement. In: Consulting Psychologists Press, 1978

[11] Hajimirza, S. N., Proulx, M. J., Izquierdo, E.: Reading Users' Minds From Their Eyes: A Method for Implicit Image Annotation. In: Multimedia, IEEE Transactions on, 14(3), 805-815, 2012.

[12] Huang, C. L., Huang, Y. M.: Fcaial expression recognition using model-based feature extraction and action parameters classification. In Journal of Visual Communication and Image Representation, 8(3), 278-290, 1997.

[13] Ishikawa, T.: Passive driver gaze tracking with active appearance models. Carnegie Mellon University, the Robotics Institute, 2004.

[14] Jesorsky, O., Kirchberg, K. J., Frischholz, R. W.: Robust face detection using the hausdorff distance. In: Audio- and video-based biometric person authentication (pp. 90-95), Springer Berlin Heidelberg, 2001.

[15] Jiao, J., Pantic, M.: Implicit image tagging via facial information. In: Proceedings of the 2nd International workshop on Social signal processing (pp. 59-64), ACM, 2010.

[16] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition, 2006, IEEE Computer Society Conference on (Vol. 2, pp. 2169-2178), IEEE, 2006.

[17] Lee, E. C., Park, K. R.: A robust gaze tracking method based on a virtual eyeball model. In: Machine Vision and Applications, 20(5), 319-337, 2009.

[18] Lim, M. Y., Aylett, R.: A new approach to emotion generation and expression. In: Proceedings of the Doctoral Consortium, The 2nd International Conference on Affective Computing and Intelligent Interaction (pp. 147-154), 2007.

[19] Marks, T. K., Hershey, J. R., Movellan, J. R.: Tracking motion, deformation and texture using conditionally Gaussian processes. In: Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(2), 348-363, 2010.

[20] Martinez Bedard, B.: Is Core Affect a Natural Kind? Philosophy Theses, 42, 2008.

[21] Milborrow, S., Morkel, J. Nicolls, F.: The muct landmarked face database. Pattern Recognition Association of South Africa, 2010.

[22] Nordstrøm, M. M., Larsen, M. Sierakowski, J. Stegmann, M. B.: The IMM face database-an annotated dataset of 240 face images. DTU Informatics, Building 321, 2004.

[23] Otsu, N,: A threshold selection method from gray-level histograms. In Automatica, 11(285-296), 23-27, 1975.

[24] Pogalin, E., Redert, A., Patras, I., Hendriks, E. A.: Gaze tracking by using factorized likelihoods particle filtering and stereo vision. In: 3D Data Processing, Visualization and Transmission, Third International Symposium on (pp. 57-64), IEEE, 2006.

[25] Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: ACM Transactions on Graphics (TOG) (Vol 23, No 3, pp. 309-314), ACM, 2004.

[26] Russell, J. A.: Core affect and the psychological construction of emotion. In: Psychological review, 110(1), 145, 2003.

[27] Sadeghi, M., Tien, G., Hamarneh, G., Atkins, M. S.: Hands-free interactive image segmentation using eyegaze. In: SPIE Medical Imaging (pp.72601H-72601H), International Society for Optics and Photonics, 2009.

[28] Salojärvi, J., Puolamäki, K., Kaski, S.: Implicit relevance feedback from eye movements. In: Artificial Neural Networks: Biological Inspirations – ICANN 2005 (pp. 513-518), Springer Berlin Heidelberg, 2005.

[29] Shan, M. K., Kuo, F. F., Chiang, M. F., Lee, S. Y.: Emotion-based music recommendation by affinity discovery from film music. In: Expert Systems with Applications, 36(4), 7666-7674, 2009.

[30] Simon, D., Craig, K. D., Gosselin, F., Belin, P., Rainville, P.: Recognition and discrimination of prototypical dynamic expressions of pain and emotions. In: Pain, 135(1-2), 55-64, 2008.

[31] Smith, C., Scott, H.: A componential approach to the meaning of facial expressions. In: The psychology of facial expression, 229, 1997.

[32] Soleymani, M., Pantic, M.: Human-centered implicit tagging: Overview and perspectives. In Systems, Man and Cybernetics (SMC), 2012, IEEE International Conference on (pp. 3304-3309), IEEE, 2012.

[33] Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. In: Affective Computing, IEEE Transactions on, 3(1), 42-55, 2012.

[34] Soleymani, M., Pantic, M., Pun, T.: Multimodal emotion recognition in response to videos. In: Affective Computing, IEEE Transactions on, 3(2), 211-223, 2012.
[35] Strupp, S., Schmitz, N. Berns, K.: Visual-based emotion detection for natural man-machine interaction. In KI 2008: Advances in Artificial Intelligence (pp. 356-363), Springer Berlin Heidelberg, 2008.
[36] Talbot, J., Xu, X.: Implementing Grabut. Bringham Young University, 2006.
[37] Terissi, L. D., Gómez, J. C.: 3D head pose and facial expression tracking using a single camera. In Journal of Universal Computer Science, 16(6), 903-920, 2010.
[38] Tkalčič, M., Odić, A., Kočir, A., Tasič, J. F.: Affective Labeling in a Content-Based Recommender System for Images. In: Multimedia, IEEE Transactions on , vol.15, no.2, pp.391,400, Feb, 2013.
[39] Viola, P. Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition, 2001, CVPR 2001, Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 1, pp. I-511), IEEE, 2001.
[40] Vinciarelli, A., Suditu, N., Pantic, M.: Implicit human-centered tagging. In: Multimedia and Expo, 2009, ICME 2009, IEEE International Conference on (pp. 1428-1431). IEEE, 2009.
[41] Vrochidis, S., Patras, I., Kompatsiaris, I. : An eye-tracking-based approach to facilitate interactive video search. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval (p. 43), ACM, 2011.
[42] Yik, M.: Studying affect among the Chinese: The circular way. In Journal of personality assessment, 91(5), 416-428, 2009.
[43] Yik, M., Russell, J. A., Steiger, J. H.: A 12-point circumplex structure of core affect. In Emotion-APA, 11(4), 705, 2011.
[44] Zeng, Z., Pantic, M., Roisman, G. I., Huang, T. S.: A survey of affect recognition methods: Audio, visual and spontaneous expressions. In: Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31(1), 39-58, 2009.
[45] Zhu, J., Yang, J.: Subpixel eye gaze tracking. In: Automatic Face and Gesture Recognition, 2002, Proceedings, Fifth IEEE International Conference on (pp. 124-129), IEEE, 2002.