AN APPLICATION FRAMEWORK FOR IMPLICIT SENTIMENT HUMAN-CENTERED TAGGING USING ATTRIBUTED AFFECT

Konstantinos C. Apostolakis, and Petros Daras

Centre for Research and Technology - Hellas
Information Technologies Institute
Thessaloniki, Greece

ABSTRACT

In this paper, a novel framework for implicit sentiment image tagging and retrieval is presented, based on the concept of attributed affect. The user's affective response is recorded and analyzed to provide an appropriate affective label, while eye gaze is monitored in order to identify a specific object depicted in the scene, which is attributed as the cause of the user's current state of core affect. Through this procedure, automatic tagging of content, as well as retrieval based on personal preferences is possible. Our experiments show that our framework successfully channels behavioral tags (in the form of affective labels) to the data tagging and retrieval loop, even when applied in the context of a cost-efficient, widely available hardware setup, that uses a single low resolution webcam mounted on a standard modern computer system.

1. INTRODUCTION

In the context of Implicit Human-Centered Tagging (IHCT), attempts are made in order to obtain user behavioral response implicitly, therefore effectively reducing user effort, in contrast to traditional tagging processes, such as textual annotation. Pantic and Vinciarelli [1] documented the challenges posed by such an endeavor; mainly concerning the need to include the observed user reactions and behavior (as well as the implicit tags themselves) to the data tagging and retrieval loop and also, the development of behavioral analyzers, capable to attain both accurate and reliable results, even when the audiovisual sensors used to obtain behavioral information are mounted on today's commercial computers. Related work on the subject such as [10] and [11] are still far from being employed in a real-life average user use case scenario, due to complexity of the sensory apparatus used to obtain data.

A major component of user behavioral response is manifested through core affect, a dimensional value constituted by a measure of valence and arousal. Core affect was defined by Russell [2] as one of the two fundamental components of a psychological framework, the other being the perception of the affective quality withheld (but not exclusively) by external stimuli. In the context of this psychological framework, any subconscious attempts to attribute changes in one's core affect state to their perceived causes define the notion of attributed affect. A stimulus identified as the cause of core affect change, will henceforth be labeled as the "Object". The latter plays a crucial role during the unfolding of an emotional episode: Attention is shifted towards, and behavior is directed at the "Object", triggering a number of high-level cognitive processes, such as primary action regulation, facial expression display and appropriate gaze shifts.

In this paper, a novel framework introducing Attributed Affect to the implicit image tagging and retrieval loop is presented, and a cost-efficient implementation using a single low resolution web camera is demonstarated in order to meet the requirements for accurate and readily available behavioral analyzers. By reverse-engineering the process of subconscious attribution, user affective response is obtained, and the "Object", a specific element depicted in the scene and obtained through image segmentation is identified by recognizing an area where the user has been focusing his/her attention on via eye gaze information. The affective quality corresponding to the obtained response is attributed to the extracted "Object", based on the assumption that the cognitive processes were triggered by the perception of the stimulus' emotional value. With the "Object" providing the retrieval query, and the affective quality used to generate an appropriate affective tag, 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 experience similar affective responses when presented with closely-related stimuli, automatic annotation of large portions of the image database is possible by allowing users to browse and look at only a limited number of images.
2. Annotation is based on user personal experience, which further addresses culture-dependant annotation problems.
3. Retrieval and recommendations can be made based on the annotated stimuli.

This work was supported by the EU funded IP project REVERIE under contract 287723

4. Our method can be applied to a multitude of hardware setups, ranging from consumer-grade hardware to more elaborate high-precision sensors, as long as there's a way to obtain gaze and affective response information.

The rest of this paper is organized as follows: Section 2 introduces the proposed framework and its components, further refined in Sections 2.1, 2.2 and 2.3. Section 3 presents the experimental results obtained through an application developed around the framework, that further support its applicability to today's commercial systems. Finally, Section 4 concludes the article with an insight on future work on the subject.

2. THE PROPOSED FRAMEWORK

2.1. Obtaining user affective response

Many methods for obtaining user affective response have been presented in the scientific literature on human-computer interaction (HCI), with each one capable of fitting to the requirements of the proposed framework. For our experimental application, focused mainly on cost-efficient behavioral analyzers, facial expression analysis in terms of Ekman's FACS Action Units (AUs) [3] was utilized for generating an appropriate affective label corresponding to the user's current facial expression display.

For the purpose of monitoring facial expressions using a single low resolution camera, several key facial features corresponding to AU muscle groups were tracked by fitting an appropriately built 68-landmark Active Shape Model (ASM) [4] onto the face area recorded on the camera frame. Key AUs whose activation is being tracked include lip corner puller and depressor (AU12 and AU15), inner and outer brow raisers (AU1 and AU2), brow lower (AU4) and jaw drop (AU26). A reference line was drawn between inner eye corners to measure landmark distance variations occurring in each frame, in reference to an offline neutral expression snapshot taken in an offline step. After AU intensity is determined by the magnitude of each landmark point displacement, valence and arousal components of user current core affect state are calculated using Equations 1 and 2:

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

$$Arousal = P - AU15 - 0.30125 \quad (2)$$

Where

$$P = \frac{1.30125}{5}(AU1 + AU2 + AU4 + AU12 + AU26) \quad (3)$$

Appropriate affective labels (tags) are obtained via this information by placing normalized values of these tuples into one of the 2D Circular Affective Space segmments, as described in the work of Yik et al [5].

2.2. Identifying the "Object"

In order to recognize the specific stimulus that triggered the affective facial expression display, the user's eye gaze has to be monitored, and a gaze point on the screen needs to be localized. Single image gaze tracking techniques have been developed to track the user's iris centre position using a single camera, and are therefore utilized within the context of our cost-efficient implementation of the proposed framework. The iris centre is located through an automatic adaptive thresholding technique, while gaze point estimation is achieved via a calibration step and linear 2D mapping, as is described in [6].

Identifying a gaze point on screen, which enables the automatic generation of a rectangular region of interest (ROI) inside the image, is only the first step towards acquiring the "Object". The GrabCut segmentation algorithm [7] is applied on the image using the ROI information as input, for segmenting the (unrelated to the emotional information) background from the relevant foreground – usually an identifiable physical object depicted in the image. The GrabCut algorithm offers the significant advantage of requiring such minimal input information, demonstrating exceptional extraction quality, and has been found by the authors to satisfyingly work as part of a solely gaze-operated image segmentation system, despite any precision limitations presented by the tracker output data.

The extracted alpha matte foreground image should depict the "Object", and is immediately attributed with the corresponding affective label currently associated with the displayed user facial expression. This information will serve as a retrieval query in order to identify other images in the database containing a depiction of this "Object" and automatically annotate them appropriately, under the assumption that the viewing of the same stimulus will convey the same emotional response when again viewed by the user (for example, identifying a certain stimulus as "disgusting" in one image, can only lead to the conclusion that seeing the same stimulus in an other image will again result in a "disgusted" user reaction).

2.3. Recognition and retrieval

For the purpose of identifying instances of the "Object" contained in other images of the database, as well as recognizing the "Object" as a retrieval request, a standard Bag of Words (BoW – commonly referred to as Bag of Features in computer vision) [8] pipeline handles object recognition requests based on the foreground image as a query. The pipeline consists of 3 stages, with region

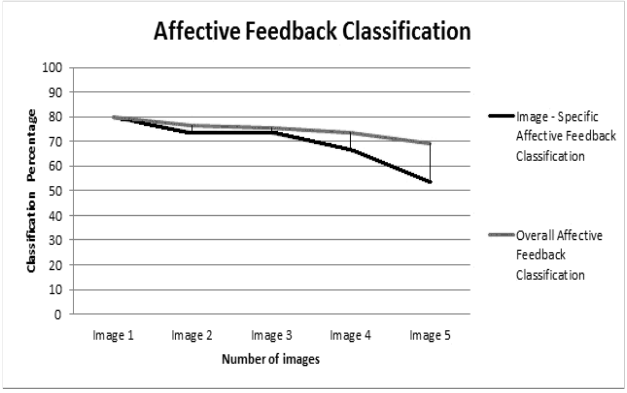

Fig. 1. Image-specific and overall affective feedback classification rates for multi-image case experiment.

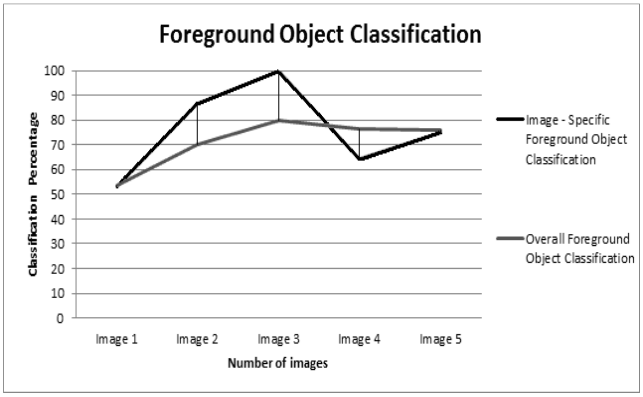

Fig. 2. Image-specific and overall foreground image classification rates for multi-image case experiment.

descriptors of the image being obtained in the first step, and projected onto a previously constructed visual vocabulary, resulting in a frequency histogram representation. The final step concerns classification of these histograms, placing the query under a certain image category, which provides the framework with the information needed to proceed with automatic annotation and retrieval.

With an emphasis on balance between accuracy and speed, obtaining image region descriptors in the first step of the BoW pipeline is achieved through Speeded Up Robust Feature (SURF) descriptors [9]. SURF is a spatial descriptor, whose responses are simple operations, like summations and subtractions, therefore achieving fast feature detection and descriptor extraction. After a number of keypoints are detected in the images, a SURF extractor can be used to compute 64-element sized descriptors out of these keypoints. A visual vocabulary is trained from a set of descriptors and kmeans clustering, resulting in the actual words, i.e. the cluster centers.

Extracting BoW image descriptors is done by first computing SURF descriptors of a given image before the nearest visual words in the vocabulary are matched to each extracted descriptor, resulting in a normalized codebook frequency histogram of vocabulary words encountered in the image (thus the i-th bin of the histogram represents the frequency of the i-th word of the vocabulary in the given image). Finally, codebook frequency histograms are classified with respect to classification speed for real-time application, using a Support Vector Machine with a Radial Basis Function (RBF) kernel.

3. EXPERIMENTAL RESULTS

In order to confirm the applicability of the proposed framework to a cost-efficient setup, thus challenging the requirements described in [1], an application was developed and evaluated for implicit tagging and retrieval use cases by 15 volunteer subjects. The application was built to utilize a single low resolution webcam for attaining user affective response and estimating gaze ROI in order to annotate images with appropriate affective tags. Both camera and machine learning modules of the application were developed with the OpenCV API[1], while tracking of the facial features required for facial expression analysis was done via ASM fitting, using the ASMLibrary SDK[2].

Experimental application runs were conducted using an in-house, self-obtained database containing 1125 images depicting the most frequently appearing distinct image categories returned by Google Images as results to the keyword “Paris”. The resulting five image categories comprising the “Paris” database included images depicting several monumental landmarks of the city (Eiffel Tower, Notre Dame, Arc de Triomphe and the Louvre Museum) as well as images of a celebrity figure with the same name.

The experimental application ran on a 3.30 GHz Intel Core i5-2500K desktop computer with a 24’’ Samsung SyncMaster 2494 display monitor and a Unibrain Fire-I 1.2 firewire 640x480 resolution webcam mounted on top of the screen. The subjects were asked to sit in front of the camera, and undergo a quick eye tracker calibration procedure using 8 points on the borders of the screen. All subjects were informed of the application scenario, a display of retrieval results based on the keyword “Paris”, and were asked to simply look at the images on display (one per category), expressing their emotions on the relevance of each image to the given keyword (displaying positive feedback on images which they would like to include to the results of a supposed future search using the same keyword, and negative feedback otherwise).

The results obtained by our experimental application showed that the application achieved an approximate 70% overall correct affective feedback classification performance (subjects’ facial expressions were correctly associated to the resulting affective labels tagged to each image) and are presented in Figure 1. Foreground object classification

[1] http://opencv.willowgarage.com/wiki/
[2] http://code.google.com/p/asmlibrary/

results presented in Figure 2, further showed an approximate 76% overall classification accuracy performance, with images belonging to the Celebrity category being 100% correctly identified from the landmark categories. These results confirm that the proposed framework successfully administers affective tags to the image tagging and retrieval loop, while also reliably attaining user behavioral response based on cost-efficient hardware widely available to users worldwide.

Several implementation choices and limitations are believed to have hindered our experimental application achieve even better results. Lack of a distinct 3-DOF head tracking module meant that facial feature tracking might have gone lost when subjects' head position significantly differed from its initial, neutral-snapshot set position. Figure 1 results showing a steady decrease in affective feedback classification, as subjects moved their heads to acknowledge the next image on display further support this conclusion. Also, examination of the foreground images showed that several images lacked the actual "Object" being depicted, causing the BoW pipeline to proceed with incorrect classification. This might be attributed to dodgy gaze point estimation, caused by camera limitations, the "Object" not being entirely contained within the rectangular ROI, or simple limitations of the GrabCut algorithm on segmenting objects containing holes (for example, the Eiffel Tower). However, further examination showed that images containing depictions of two different object categories (for example, Arc de Triomphe and Eiffel Tower), correctly identified which object was being looked at and proceeded with tagging the correct category of images, further reinforcing applicability of the framework to tagging and retrieval use cases.

4. CONCLUSIONS AND FUTURE WORK

In this paper, a novel content based image tagging and retrieval framework was presented, based on the concept of attributed affect. The latter serves as a definition of emotional awareness, serving as a main route to the affective quality of distinct stimuli, that influence the experience of core affect during the unfolding of an emotional episode. The framework takes advantage of this attribution process, and by identifying user affective response and focus of attention, provides a means to distinctly characterize specific objects depicted in an image and allow for automatic annotation of multiple images containing these stimuli, as well as a basis for retrieval based on each user's emotional preferences. An experimental application was developed around this framework, and its evaluation attested to our framework's applicability to the data tagging and retrieval loop, even on low-budget, commonly available hardware, mounted on today's computer systems.

Our future endeavors concern the benefits and application of our framework to a number of actively open research fields on data tagging and retrieval applications, such as content-based recommender systems, tagging and retrieval on complex image scenes and automatic annotation and classification of video sequences through audience reaction monitoring (therefore making emotional video segments searchable). Furthermore, we plan to apply our framework towards object and event recognition, emotional classification and display inside immersive virtual environments.

5. REFERENCES

[1] M. Pantic and A. Vinciarelli, "Implicit human-centered tagging", *IEEE Signal Process Mag* **26**, 173 – 180, 2009.

[2] J. A. Rusell, "Emotion, core affect, and psychological construction", *Cognition and Emotion* **23(7)**, 1259 – 1283, 2009.

[3] P. Ekman and W. V. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement," *Consulting Psychologists Press*, 1978.

[4] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, "Active shape models – their training and application," *Computer Vision and Image Understanding* **61**, 38 – 59, 1995.

[5] M. Yik, J. A. Russell and J. H. Steiger, "A 12-Point Circumplex Structure of Core Affect", *Emotion* **11**, 705 – 731, 2011.

[6] J. Zhu and J. Yang, "Subpixel eye gaze tracking," *Proceedings of the fifth IEEE International Conference on Automatic Face Gesture Recognition*, 124 – 129, 2002.

[7] C. Rother, V. Kolmogorov and A. Blake, "GrabCut: interactive foreground extraction using iterated graph cuts," *Computer* **23** (3), 301 – 314, ASSOC COMPUTING MACHINERY, 2004.

[8] D. D. Lewis, C. Nédellec and C. Rouveirol, "Naïve Bayes at Forty: The Independence Assumption in Information Retrieval," *Proceedings of ECML-98, 10th European Conference on Machine Learning*, Chemnitz, DE: Springer Verlag, Heidelberg, DE. Pp. 4 – 15, 1998.

[9] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, "SURF: Speeded Up Robust Features," *Computer Vision and Image Understanding (CVIU)*, Vol 100, No. 3, pp. 346 – 359, 2008.

[10] J. Jiao and M. Pantic, "Implicit image tagging via facial information," in *Proceedings of the 2nd International Workshop on Social Signal Processing*, ACM, pp. 59-64, 2010.

[11] S. Koelstra and I. Patras, "Fusion of facial expressions and EEGfor implicit affective tagging," *Image and Vision Computing*, 2012