Human Behavior Sensing for Tag Relevance Assessment

Mohammad Soleymani
Dept. of Computing
Imperial College London, UK
m.soleymani@imperial.ac.uk

Sebastian Kaltwang
Dept. of Computing
Imperial College London, UK
sk2608@imperial.ac.uk

Maja Pantic
Dept. of Computing/ EEMCS
Imperial College London, UK/
Univ. Twente, Netherlands
m.pantic@imperial.ac.uk

ABSTRACT

Users react differently to non-relevant and relevant tags associated with content. These spontaneous reactions can be used for labeling large multimedia databases. We present a method to assess tag relevance to images using the non-verbal bodily responses, namely, electroencephalogram (EEG), facial expressions, and eye gaze. We conducted experiments in which 28 images were shown to 28 subjects once with correct and another time with incorrect tags. The goal of our system is to detect the responses to non-relevant tags and consequently filter them out. Therefore, we trained classifiers to detect the tag relevance from bodily responses. We evaluated the performance of our system using a subject independent approach. The precision at top 5% and top 10% detections were calculated and results of different modalities and different classifiers were compared. The results show that eye gaze outperforms the other modalities in tag relevance detection both overall and for top ranked results.

Categories and Subject Descriptors

H3.3 [**Information Search and Retrieval**]: Information filtering, Selection process

Keywords

implicit tagging, eye gaze, EEG, facial expressions

1. INTRODUCTION

We are witnessing a rapid growth in the number of images and videos captured by users. The proliferation of handheld devices with built-in cameras is the main contributor of this rapid growth. This rapidly growing content is in need of effective indexing to be browsable and reusable. Tags are any form of metadata that can be used to index multimedia content to facilitate its finding and re-finding. In contrast to classic tagging schemes where the users' direct input is mandatory, human-centered implicit tagging was proposed [9] to gather tags and annotations without any effort from users. The main idea behind this passive tagging strategy is to use users' spontaneous reactions to a given content to identify tags. The resulting tags are called "implicit" since there is no need for users' direct input, and reactions to multimedia are displayed spontaneously [9].

Implicit tagging has recently attracted attention of the research community [13]. Implicit tagging has been used for image annotation, video highlight detection, topical relevance detection and retrieval result re-ranking. The existing literature can be divided into two categories, one dealing with using emotional reactions to tag the content with the expressed emotion, e.g., laughter detection for hilarity [10], and the second group of studies using the spontaneous reactions for information retrieval or search results, e.g., eye gaze for relevance feedback [4].

Users respond differently to the expected (i.e. relevant) and mismatching (i.e. non-relevant) tags. Koelstra et al. [6] found significant differences between N400 Evoked Related Potential (ERP) in Electroencephalogram responses between relevant and non-relevant tags displayed on short videos. Facial expression and eye gaze were used to detect users' agreement or disagreement with the displayed tags on images [5, 12]. The results showed that not all the subjects in the experiment were expressing their agreement or disagreement on their faces and their eye gaze were more informative for agreement assessment. Eye gaze responses have been also used to detect interest for image annotation [3], relevance judgment [11], interactive video search [16], and search personalization [1].

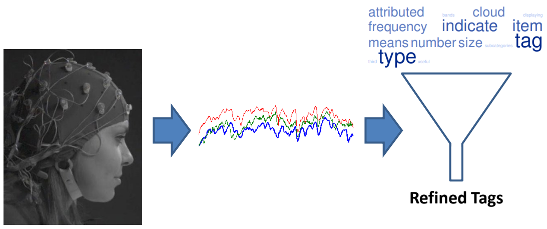

Figure 1: In our implicit tagging scenario, the information sensed from different sensors can be used to refine a set of tags assigned to images.

We propose a tag relevance assessment method to detect users' agreement with the displayed tag with a given content, i.e., images. Such a system can be used to filter out non-relevant tags from a noisy set of tags. Tags that are detected by an imperfect content based tagging system or user generated tags often contain non-relevant tags which are in-

MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502172.

correctly assigned to content. In our proposed method, by sensing the non-verbal behavioral response of a user, our system will be able to better identify the non-relevant tags and discard them. We studied the performance of a multimodal approach using three different modalities: EEG, facial expressions and eye gaze. A schematic representation of our proposed method is shown in Figure 1.

2. APPARATUS AND DATA COLLECTION

The experimental data was collected from 28 healthy volunteers, comprising 12 male and 16 female between 19 to 40 years old. The subjects had normal or corrected to normal vision. The eye gaze response was collected at 60 Hz using a Tobii X120 Eye gaze tracker[1]. The experiment was controlled by the Tobii studio software. EEG signals were recorded from 32 active electrodes on 10-20 international system using a Biosemi Active II system. The frontal facial video was captured using an Allied Vision Stingray F-046B monochrome camera with the resolution of 780 × 580 pixels at 60 frames per second. The experiments were conducted in a room with controlled temperature and illumination. The synchronization method, hardware setup and the database details are given in [12]. MAHNOB-HCI is a publicly available database for multimedia implicit tagging[2].

During the experiment, 28 images depicting human actions (e.g. handshake) were subsequently shown on their own and accompanied by a word tag that is either relevant or non-relevant to the shown action. Images were downloaded from Flickr[3] and were cropped and resized to 1280×695 pixels to be displayed on a display size of $51.9 \times 32.45 cm$ with a resolution of 1280 × 800 pixels. The space under and above the image was filled with black pixels. The tags were overlaid under the image (see Figure 2). For each image a correct and an incorrect tag was displayed in the total of 54 trials in random order. For each trial, the following procedure was taken. First, the untagged images were displayed for 5 seconds. This allowed the subject to get to know the content of the image. Second, the same image was displayed with a tag for 5 seconds. The subjects' behavior in this period contained their reaction to the displayed tag. Third, a question was displayed on the screen to ask whether the subject agreed with the suggested tag. Agreement or disagreement was expressed by pressing a green button for relevance or a red button for non-relevance feedback, respectively. The length of each trial was about 11 seconds. In this study, the trials in which the subjects' responses contradict the true tag relevance were discarded as confusing examples, e.g., a trial in which a subject agreed with a non-relevant tag was discarded.

3. DATA ANALYSIS

3.1 Analysis of Eye Gaze

A set of features is extracted from the different signals provided by the eye gaze tracker. These signals include gaze fixation, pupil dilation, scan path and eye gaze coordinates. Gaze fixations are the coordinates of the points on the display on which the eye gaze stayed fixed for a certain period of time. Each fixation is composed of its duration as well as the two-dimensional coordinates of the projection of the eye gaze on the screen. An example of an eye gaze pattern and fixation points on an image is shown in Figure 2. The scan path is the eye gaze trajectory in transition from one fixation to another. The image zone and the tag zone were defined on the screen based on the way the images were positioned on the display and the image size, which was constant. The features extracted from eye gaze are listed in Table 1.

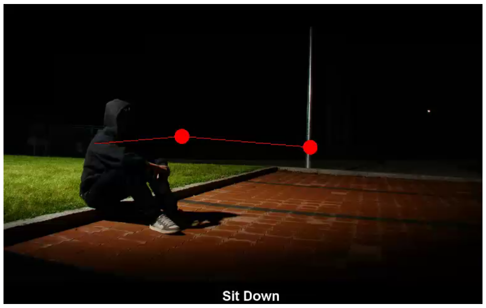

Figure 2: Example image depicting a human action including a relevant tag ('Sit Down') as shown to the subjects. Part of the recorded eye gaze fixation and scan path of one subject is overlaid in red.

Table 1: List of 26 features extracted from eye gaze data for implicit tagging. The number of features extracted from each signal is indicated in brackets. Average (avg.) and standard deviation (std.) are abbreviated.

Signal	Extracted features
Pupil (2)	Avg. and std. of the dilation
Eye blink (2)	Avg. blink duration and blinking rate
Distance (1)	Range of viewer's distance to the screen
Scanpath (5)	Number of transitions between the tag zone and the image zone, avg. scan path length, std. of the scan path length, total length of the scan path
Saccades (2)	Avg. of standard deviations of eye gaze movements during each fixation period in horizontal and vertical axis
Fixation (6)	Avg. fixation duration in the tag zone, avg. fixation duration in the image zone, max. of the fixation duration in the tag zone, max. of the fixation duration in the image zone, number of fixations in the tag zone divided by the number of fixations in the image zone and number of fixations in the tag zone
Eye gaze (8)	Avg., std., skewness and kurtosis of horizontal and vertical gaze coordinates

3.2 Analysis of EEG

EEG signals were originally recorded with a 1024Hz sampling rate. The unwanted artifacts, trend and noise were reduced prior to extracting the features from EEG data by pre-processing the signals. Biosemi active electrodes record EEG signals referenced to common mode sense electrode (CMS) as a part of its feedback loop. In order to gain the

[1] http://www.tobii.com
[2] http://mahnob-db.eu/hci-tagging/
[3] http://www.flickr.com

full common-mode rejection ratio (CMRR) at 50Hz, EEG signals should be re-referenced to another reference. EEG signals were thus re-referenced to the average reference to maximize signal to noise ratio. EEG drift was removed by subtracting the moving averaged signal with a 5 seconds window. The noise reduction was done by applying a low-pass filter with the cut-off frequency of 10Hz, since the ERP responses are low frequency [7].

We expected the ERP responses to appear in 400ms to 600ms after showing the tag. Therefore, the EEG signals of the one second period after displaying overlaid tags under the images were downsampled 16 times and used as features. As a result, we had $32 \times 16 = 512$ features for every trial.

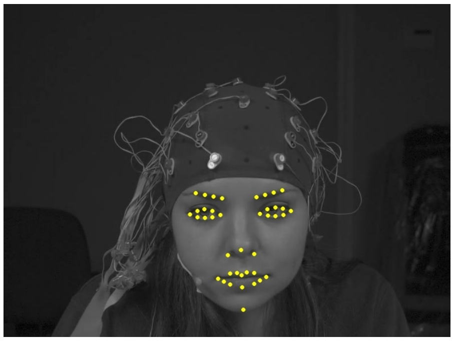

Figure 3: An example of the recorded camera view including tracked facial points.

3.3 Analysis of facial expressions

An active appearance model face tracker was employed to track 40 points [8] (see Figure 3). The facial points were extracted after registering the face to a normalized face and correcting the head pose. A reference point was generated by averaging the inner corners of eyes and points on the subjects' nose which assumed to be static. The distances of 33 point including eyebrows, eyes, lip and iris to the reference point were calculated and averaged to be used as features.

3.4 Classification

For this study, we conducted subject independent tag relevance detection. In this subject independent approach, we perform a leave-one-subject-out cross-validation, i.e. all trials of one subject are used for testing and all trials of the remaining subjects are used for training. This is repeated until each subject has been used for testing.

We used a Relevance Vector Machine (RVM) [15], a Support Vector Machine (SVM) [2] and Linear Discriminant Analysis (LDA) as classifiers. Both, the RVM and SVM are linear classifiers in the kernel space. The kernel space is defined by the training dataset and only a sparse subset is used for classification, the so called Relevance Vectors or Support Vectors. However, their optimization technique is different: the SVM formulates the problem within a max margin framework, i.e., the algorithm maximizes the margin around the hyperplane that discriminates between the classes. The RVM formulation is fully probabilistic, i.e., it defines a model fitting probability for the training data and a prior over the model parameters. The algorithm then proceeds by optimizing the posterior probability of the parameters given the data.

For RVM and SVM, we use a Gaussian radial basis function kernel with length scale parameter γ. Additionally, the SVM has the soft-margin cost parameter C. Both parameters are optimized within each cross-validation fold by a grid-search. We first calculate the mean value m of the pairwise kernel functions between all training data points and then start the grid-search within the interval $[m/2; 2m]$ for γ and $[1; 10]$ for C. The interval is searched with a logarithmic step size and it is extended if the optimal value is found at the border. The classifier is evaluated for each of the steps by a 2-fold cross-validation.

The LDA classifier is applied after reducing the dimensionality of features using a Principal Component Analysis and preserving 95% of the energy.

4. EXPERIMENTAL RESULTS

The goal of our proposed method is to filter out non-relevant tags. Therefore, we report the precision of the top 5% and 10% samples detected as non-relevant in Table 2. From the precision @5% and @10% SVM has a slight advantage for eye gaze and LDA for the other modalities. Facial expression analysis do not yield very convincing results. This can be due to the very subtle responses which were also highly varying between subjects. EEG signals have also a strong person specific component and ERP responses are not easily extractable from one trial. If we want to take advantage of EEG for such methods, the tag will have to be flashed for multiple times in order to get a sufficient level of signal to noise ratio. Eye gaze results are significantly better than the other modalities. Thus, after fusion of these modalities at decision level we could not achieve any better performance than the best single modality. Precision and recall curves of the best modality, eye gaze, is shown in Figure 4. The precision and recall results were calculated in each iteration of cross validation and then averaged to present the general performance of the system. Looking at the precision and recall curves, SVM performs the best to detect the top ranked non-relevant results and therefor is a good candidate for a filtering system that removes the top ranked non-relevant tags.

We examined the correlation between the labels and the different eye gaze features as indicators for their relevance. The fixation time in the tag zone was in average longer for non-relevant tags which means the participants spent more time looking at the labels. The vertical position of the eye gaze was more towards the bottom of the screen for non-relevant tags which is as a result of spending more time in the tag zone.

To compare our system to the previously published results, we calculated the classification rate and F1 scores. The best subject independent classification rate was obtained using only eye gaze whose classification rate is 59.5% with the average F1 score of 0.59. This is superior to the average subject dependent results reported in [5, 12, 14] and in the same line with the subject dependent results reported in [3] using eye gaze, the best F1 score reported was 0.6. It is worth noting that a subject independent approach is more useful in practice due to the higher chance of generalizing over a population.

5. CONCLUSIONS

We presented a method based for filtering non-relevant tags on images. We obtained promising results and our find-

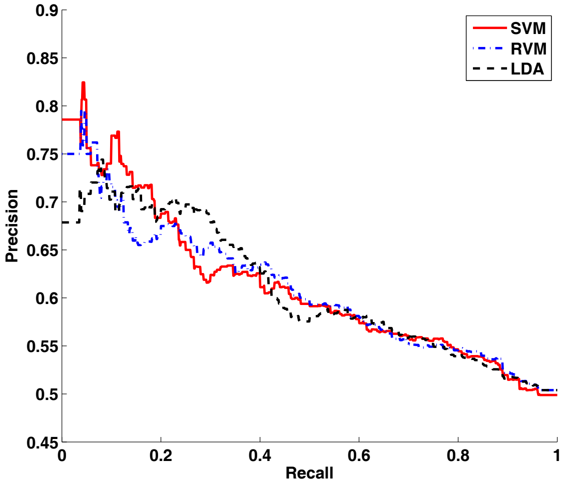

Figure 4: Recall and Precision curves. The RVM, SVM and LDA+PCA results on eye gaze are shown in different colors.

Table 2: The precision at top 5% and top 10% non-relevant detection for different modalities, namely, EEG, Eye Gaze (EG), Facial Expressions (FE) and different classifiers.

	RVM		SVM		LDA+PCA	
Metric	@5%	@10%	@5%	@10%	@5%	@10%
EEG	0.50	0.56	0.35	0.42	0.57	0.59
EG	0.73	0.70	**0.76**	**0.73**	0.74	0.73
FE	0.51	0.52	0.49	0.48	0.55	0.55

ings can pave the way for the future studies on this topic with possibly larger data collections. Although the overall detection rate is far from ideal, our system is able to detect more accurately the top ranked non-relevant results. This enables us to filter out non-relevant tags with a high confidence. We can repeat this with more than one subject to filter out more of the noisy and incorrect tags with higher confidence. Studying three different modalities, EEG, facial expressions and eye gaze, we found that eye gaze was the most informative and generalizable channel of information for such applications. Given the conclusions, the future studies can focus on collecting and analyzing larger datasets from the most informative modality, i.e., eye gaze.

6. ACKNOWLEDGMENT

The work of Soleymani is supported by the European Research Council under the FP7 Marie Curie Intra-European Fellowship: Emotional continuous tagging using spontaneous behavior (EmoTag). The work of Kaltwang is partially funded by EPSRC grant EP/H016988/1 (EmoPain). The work of Pantic is supported in part by ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The authors want to thank B. Martinez and J. Orozco for the discussions and J. Doboš and J. Lichtenauer for their contributions to the data collection.

7. REFERENCES

[1] G. Buscher, L. van Elst, and A. Dengel. Segment-level display time as implicit feedback: a comparison to eye tracking. In *ACM SIGIR*, SIGIR '09, pages 67–74, 2009.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Systems and Technology*, 2:1–27, 2011.

[3] S. Hajimirza, M. Proulx, and E. Izquierdo. Reading users' minds from their eyes: A method for implicit image annotation. *Multimedia, IEEE Trans.*, 14(3):805 –815, june 2012.

[4] D. Hardoon and K. Pasupa. Image ranking with implicit feedback from eye movements. In *Symp. on Eye-Tracking Research & Applications*, ETRA '10, pages 291–298. ACM, 2010.

[5] J. Jiao and M. Pantic. Implicit image tagging via facial information. In *Int'l Workshop on Social signal processing*, pages 59–64. ACM, 2010.

[6] S. Koelstra, C. Muhl, and I. Patras. EEG analysis for implicit tagging of video data. In *Int. Conf. on Affective Computing and Intelligent Interaction.*, pages 1–6. IEEE, 2009.

[7] F. Lotte and C. Guan. Learning from other subjects helps reducing brain-computer interface calibration time. In *ICASSP 2010*, pages 614–617. IEEE, 2010.

[8] J. Orozco, O. Rudovic, J. Gonzàlez, and M. Pantic. Hierarchical On-line Appearance-Based Tracking for 3D Head Pose, Eyebrows, Lips, Eyelids and Irises. *Image and Vision Computing*, 31(4):322–340, 2013.

[9] M. Pantic and A. Vinciarelli. Implicit human-centered tagging. *IEEE Signal Processing Magazine*, 26(6):173–180, November 2009.

[10] S. Petridis and M. Pantic. Is this joke really funny? judging the mirth by audiovisual laughter analysis. In *IEEE ICME*, pages 1444 –1447, 2009.

[11] J. Salojärvi, K. Puolamäki, and S. Kaski. Implicit relevance feedback from eye movements. In W. Duch et al., editors, *ICANN 2005*, volume 3696 of *Lecture Notes in Computer Science*, pages 513–518. Springer Berlin / Heidelberg, 2005.

[12] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affective Computing*, 3:42–55, 2012.

[13] M. Soleymani and M. Pantic. Human-centered implicit tagging: Overview and perspectives. In *IEEE SMC*, pages 3304–3309, 2012.

[14] M. Soleymani and M. Pantic. Multimedia Implicit Tagging using EEG Signals. In *IEEE ICME*, 2013. in press.

[15] M. Tipping and A. C. Faul. Fast marginal likelihood maximisation for sparse Bayesian models. In *Proc. 9th Int'l Workshop on Artificial Intelligence and Statistics*, pages 1–13, 2003.

[16] S. Vrochidis, I. Patras, and I. Kompatsiaris. An eye-tracking-based approach to facilitate interactive video search. In *ACM Int. Conf. on Multimedia Retrieval*, ICMR '11, pages 43:1–43:8. ACM, 2011.