# Affective video summarization and story board generation using Pupillary dilation and Eye gaze

Harish Katti, Karthik Yadati, Mohan Kankanhalli, Chua Tat-Seng
School Of Computing, National University of Singapore
harishk, nyadati, mohan, chuats@comp.nus.edu.sg

***Abstract*—We propose a semi-automated, eye-gaze based method for affective analysis of videos. Pupillary Dilation (PD) is introduced as a valuable behavioural signal for assessment of subject arousal and engagement. We use PD information for computationally inexpensive, arousal based composition of video summaries and descriptive story-boards. Video summarization and story-board generation is done offline, subsequent to a subject viewing the video. The method also includes novel eye-gaze analysis and fusion with content based features to discover affective segments of videos and Regions of interest (ROIs) contained therein. Effectiveness of the framework is evaluated using experiments over a diverse set of clips, significant pool of subjects and comparison with a fully automated state-of-art affective video summarization algorithm. Acquisition and analysis of PD information is demonstrated and used as a proxy for human visual attention and arousal based video summarization and story-board generation.**

**An important contribution is to demonstrate usefulness of PD information in identifying affective video segments with abstract semantics or affective elements of discourse and story-telling, that are likely to be missed by automated methods. Another contribution is the use of eye-fixations in the close temporal proximity of PD based events for key frame extraction and subsequent story board generation. We also show how PD based video summarization can to generate either a personalized video summary or to represent a consensus over affective preferences of a larger group or community.**

## I. Introduction

Affective content in videos is an important, yet relatively unexplored aspect of videos. Though a large body of literature addresses semantics oriented video summarization, annotation and indexing. It remains challenging to define and represent affective elements in videos. Video content can be affective due to a diversity of visual elements (a) low level, local colour information relating to hue, saturation and intensity (b) global information that can convey mood or emotion (c) abstract scene semantics due to concepts such as faces, people and their interaction (d) discourse related elements in the video that are spread over a large time window. Intuitively, one can see that these visual elements also present an increasing order of cognitive complexity for humans and probably even computational complexity for affective video analysis methods. It is not surprising then, that even in the nascent field of affective analysis, research has gradually addressed problems beginning from (a) and moving gradually towards modelling (d) discourse and story-telling.

In this paper, we propose a novel method to use human Pupillary dilation response for identification of arousing video segments for summarization. We also demonstrate affective story board generation by identification of arousing key-frames using PD information and eye-fixations. Our current method works offline and subsequent to a video viewing session, and we show that such summaries are found satisfactory by (a) the same viewer (b) other viewers (c) against fully automated state-of-art [34].

Commonly encountered videos such as movies and personal videos have significant affective elements due to the presence of people, their interaction and emotional responses to different situations, aural components relating to music and speech can also contribute to affect [24]. Segments of such videos can represent a variety of emotions that can be categorized in many ways. A popular set of canonical emotions in psychology research is that by Ekman [7], *negative* (angry, sad, disgust), *positive* (happy) and *neutral* (surprise, lack of emotion). These emotions are clearly manifested on human faces and are found to be consistent across gender, race and most cultures [7]. An alternative representation is that of the continuous space spanned by attributes of *pleasure/valence* (degree of pleasantness or unpleasantness), *arousal* (degree of excitement or activation) in the Circumplex model [28].

There is a difference of opinion amongst the affective research community regarding the appropriate representation for a computational model and both discrete labels similar to Ekman's model [7] and continuous ratings similar to the Circumplex model [28] have been chosen in existing literature for different reasons. We observe that the choice of representation for affective video analysis has usually been determined by the mathematical model employed for learning and classifying emotional states. For example, discrete representations have been used with hidden markov models (HMMs) [33] and more recently [34]. On the other hand continuous representations have been used with bayesian networks (BN) [32], variants such as dynamic bayesian networks (DBN) [3], regression models such as support vector regression (SVR) [35] and by designing functions to represent a continuum of valence and arousal [10]. We would like to point out here that past [28] and more recent work [9] in psychology and neuroscience research has indicated and refined the correspondence between such discrete and continuous representations.

We feel this equivalence between continuous valence-arousal (VA) based representation and that using labels for discrete emotional states is important for affective analysis of videos. VA ground truth is usually obtained by asking human subjects

to annotate video or image content with continuous scores [21]. Similarly, labels can be sought for *happy*, *sad* and other emotional states [7]. These methods are more convenient and non-intrusive than physiological recordings like heart-rate, skin-conductivity, etc [19]. Two questions become relevant due to these approaches,

- How can one map the discrete and continuous ratings of emotional states ?
- Do such human annotated ratings correlate well with the actual spontaneous response of an individual ?

We find that both questions have satisfactory and encouraging answers. Studies have independently rated the same set of stimuli with both continuous rating and labeling with discrete emotions. An example is [9], where the faces from Ekman's dataset of canonical facial expressions was rated by subjects on the VA plane. Mapping between two representations shows good correlation and is also illustrated in Figure 1. 10 subjects were asked to indicate emotional rating on a 2 dimensional VA grid after being shown expressive faces exhibiting emotions such as *happy*, *sad*, etc. Good agreement is observed across subjects for VA values assigned to each emotion as can be seen by the small standard deviation bars overlaid onto manually annotated VA ratings (diamond). This address the first question. The second question is also addressed by this study as can be seen by the good, correlation between spontaneous emotional response in brain regions (stars) and manually annotated VA scores (diamonds). Though the overlap is not exact, the distribution of manually assigned VA scores and that of brain activity in specific brain regions [9] is quite similar. Brain activity was recorded in this study using BOLD fMRI based measurements.

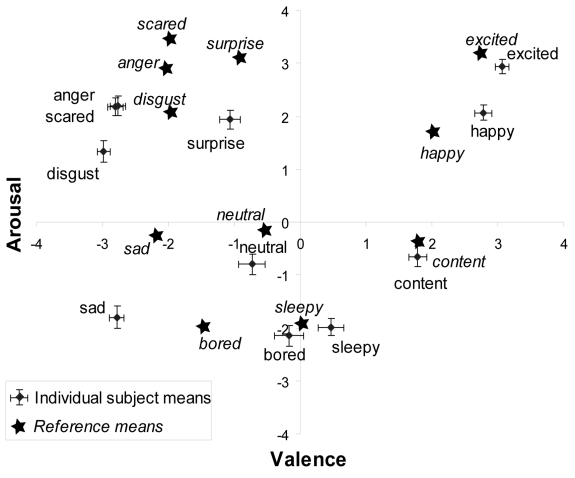



Fig. 1. Conscious, manually annotated VA ratings (diamonds) given to affective faces from the Ekman dataset show good consistency across 10 subjects. Furthermore, individual ratings (diamonds) on the VA plane by humans, correlates well with spontaneous brain activity (stars). Results from [9]

Video summarization has been attempted by three dominant approaches using video content alone, using metadata such as text transcripts and hybrid systems that use a mix of both approaches. In this paper, we employ a hybrid approach and combining content analysis and PD information as proxy for subject arousal and perform video segmentation. Affective video summaries can be multi functional and can be used as preview clips, indexing and browsing. Summaries can be useful for bandwidth constrained scenarios where users might want to get an overview of the content prior to initiating a streaming session. Commercial content such as movies, music videos, news programs and documentaries have a possibility of meta-data created during production (multilingual captioning) or by general public (open subtitles). Personal video archives are also an important and ever growing corpus of content and bring forth the need for personalization of content to suit individual preferences. Personalized summaries from amateur and home videos is an interesting and challenging direction of research. In contrast to commercial content, there is usually little or no accompanying metadata generated along with such videos. On the other hand, they offer possibilities of annotation using social networks [23] and the presence of limited number of people and activities.

Behavioural signals are a rich and readily available window into human cognition. Recent technological developments have made it easy and cost-effective to record such data on-the-fly, as a user views video content. In this paper, we focus on eye-gaze as a non invasively recorded proxy for human visual attention and accompanying PD signal as a proxy for subject's state of arousal. The cartoon in Figure 2 illustrates a typical eye-tracking and pupillary dilation measurement setup.

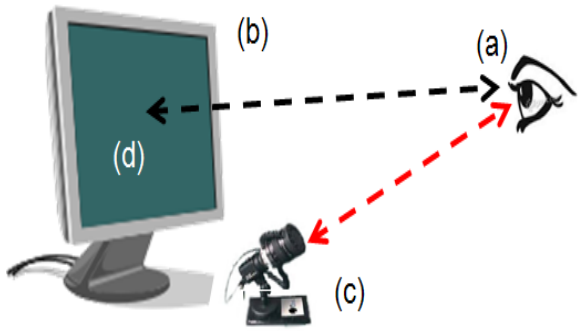



Fig. 2. Illustration of the interactive setup. A video stimuli viewing session in progress, (a) The subject looks at visual input. (b) The display screen. (d) The on-screen rendering of video frames. (c) Desktop based eye-tracker is used to establish a mapping between images of the subject's eye and the on-screen location being looked upon, while viewing the video.

Examples are recent laptop prototypes with in-built eye-tracking hardware, developed by Tobii and Lenovo. It is also possible to harness existing cameras on laptops and mobile devices to track the subjects visual attention. Opensource systems such as Opengazer [2] have been used for research and development using eye-tracking. An example is the eye-fixation inference in [8]. New generation of mobile devices like the new ipad®devices from Apple©could soon be equipped with multiple front-facing cameras as indicated by the recent patent [18] by Apple©to adjust the displayed

content according to a user's face position and viewing perspective.

Pupillary response or dilation of the pupil is a physiological response that varies the size of the pupil of the eye via the iris dilator muscle. The structure of the eye, location of the pupil along with major and minor elliptical axes are shown in Figure 3. An important function is to control the amount of light entering the eye help accommodate to different lighting conditions. Pupillary dilation (PD) refers to the increase in size of pupils, measurement of PD is also called Pupilometry. PD measurement is built into most current eye trackers. PD is typically recorded in terms of the major and minor axis corresponding to the elliptical pupil opening.

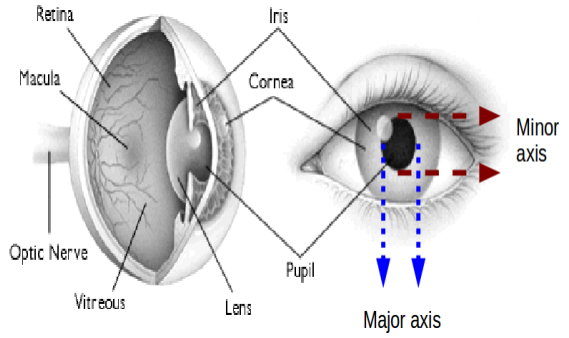



Fig. 3. Basic structure of the eye and location of the pupil. Pupillary dilation (PD) is the increase in the pupil size, it causes greater amount of light to fall on the retina. PD is also associated with emotional responses. Image copyright reserved by http://t2.gstatic.com.

The important contributions of our paper are,

- A novel method to use Pupillary dilation for identification of arousing segments of video for summarization.
- Story board generation by identification of arousing key-frames using PD information and eye-fixations.

It is important to note that our system needs a human in the loop, affective preferences learnt from such interaction can then be used for personalized summarization or a consensus taken for community preference. This is akin the use of query logs from past users in search-engine technology. Search engines use past user interaction to improve not just individual personalization, but also incrementally generalise to all users of the system. We anticipate that attention capture and affective interaction will be seamlessly integrated into media enabled devices. This scenario sets the context of our work and we explain our method and its capability in remainder of this paper.

## II. Related work and Background

### A. Human visual attention (HVA) and eye-tracking

HVA is an important strategy to humans to focus visual processing resources to select regions in the field-of-view. Eye movements help to bring an object of interest to the central region of the visual field, capable of detail processing. Eye *fixations* are defined as temporal events where eye-gaze remains restricted to within $1^o$ visual angle for 100 milliseconds or more. Eye movements leading from one fixation to another are termed as *saccades*. Most visual information assimilation happens during *fixations* and does not happen during *saccades*. Attention has been shown to be driven by visual conspicuity, popularly termed as visual saliency. Popular visual saliency models incorporate bottom-up cues derived from local properties such as intensity and color variations [15] or motion [4] and top-down cues like scene semantics involving key objects and their interactions [6]. Eye-gaze has been explored as a valuable signal to understand how humans understand static and dynamic visual scenes. In [16], authors explore how eye-gaze based saliency can be used to improve upon automated methods for saliency estimation. In [26], eye-gaze to model how human's prioritize visual concepts in images; eye-gaze statistics are fused with visual concepts to obtain a *world-model*. The relationship between eye-gaze and image semantics has been demonstrated in image caption localisation using eye-gaze in [26] and to guide object detectors find key concepts in images [17]. Eye-gaze has also been used for identifying foreground regions of interest in images using eye gaze [27]. It has been used for image cropping in [29] and has been suggested for use in video retargeting in [31].

### B. Eye-gaze acquisition and analysis

Current acquisition methods use commercial eye-trackers, which establish a correspondence between relative positions of the subject's eye (iris, pupil, etc) as the subject looks at different on-screen locations on the display. This also necessitates a calibration step where the eye configurations specific to pre-determined on-screen locations are stored. Basic statistical analysis of eye fixations has been explored in psychology and more recently in computer science literature [26]. Novel algorithms for eye-gaze clustering have been proposed in [30] and more recently in [17]. Low cost eye-tracking has been explored using ordinary web cams in the opengazer system [2] and the ITU gazetracker [1]. Video content is dynamic and also has editing artifacts such as scene changes along with camera movements, which in turn make inference of ROIs in videos difficult. This is made more difficult in interactive scenarios, which require online processing of eye-gaze data. Our frame work is one of the first to address this problem to the best of our knowledge.

### C. Video summarization and story-boards

Summarization of videos has a long history in the form of movie trailers, news flashes, etc. It is a well researched area and there have been a variety of automated and hybrid approaches [22]. This paper has a *hybrid* approach to arousal based summarization, as information external to video is being used to identify important segments. Closest in spirit to our work is that of [10], where the VA theory from psychology is employed in a framework to identify key segments of videos. We compare our affective summarization results to state-of-art in affective video summarization [34]. The authors generate personalized affective summaries from videos by

labeling segments of video with *neutral* and *happy* emotional labels. A recent work using eye fixations and facial expressions for interactive video summarization is that of [25], though similar in spirit, they rely primarily on facial expression analysis to compute emotional states of the viewer. We find an good benchmark in [34] as it is a fully automated affective analysis method and the choice of emotional states is bi-valued (neutral/positive) valence as in our case. Furthermore, we find it is more meaningful to compare our hybrid method against fully automated state-of-art.
For story-board generation, we find a close parallel in [14]. Video sequences are processed to perform speaker identification and dialogue localisation and given cartoon like effect. Key frames are then assembled into standard comic layouts. In this paper we focus on utilizing Pupillary Dilation (PD) information to extract interesting clips for summarization and key frames for story board generation.

### *D. Pupillary dilation*

The earliest reported work relating to pupil dilation goes back to Heinrich's work in 1896 [11] where he correlated PD with change in subject's attention to visual stimuli. Significant work relating to the role of PD can be found from 1960's with Eckhard Hess's work correlating PD to interest [12] and emotions such as disgust and sexual arousal [13]. More recently, PD has shown to be relevant in a visual task involving detection of errors in number sequences presented visually to subjects [20] and more recently to presentation of affective images [5]. PD values change in correspondence to multiple cues including affect, cognitive load and changes in light intensity. In video analytics such as ours, affective cues and cognitive load typically indicate arousal. Light intensity changes on the other hand can be countered by detecting and discarding PD events corresponding to sudden intensity changes in the video stream such as those on shot boundaries.

### *E. Our method for summarization and story-board generation*

We explore subject arousal as an indicator of engagement and surprise and use it to extract video segments and representative key frames to summaries video. Arousal has been shown to be a valuable cue for this purpose [10]. The important steps in our method are explained in this section.

Steps for PD based arousal event detection,

1) Record eye-movements and PD information by eye-tracking.
2) Compute elliptical pupillary dilation area $\pi \times minor_{axis} \times major_{axis} \times$.
3) Pre-process raw PD signal using smoothening over a temporal window $t$ and normalizing in the range [0,1].
4) Compute standard deviation $\sigma$ and mean $\mu$ over pre-processed PD signal.
5) Identify time segments PD deviates from mean by more than $k \times \sigma$.
6) Detect shot boundaries and reject corresponding PD peaks.

Steps for story board detection,

1) Identify first fixation following each PD arousal peak.
2) Mark corresponding video frames as a key-frames.
3) Concatenate key-frames to a story board sequence.

PD is computed as the area of the elliptical pupillary opening $\pi \times minor_{axis} \times major_{axis} \times$. Sudden deviations in PD are normally indicators of arousal and subject engagement and are shown by the blue trace in Figure 4 (a). The raw PD signal as shown by the red trace in Figure 4 (a), is smoothed using a moving average of few seconds (green trace in Figure 4 (a)) to accommodate for the fact that arousal variations are slow events occurring over a span of several seconds. The PD standard deviation value $\sigma$ computed and video segments corresponding to a $k \times \sigma$ deviation from mean PD value are identified. The value of $k$ is variable and can be set based on user's profile or in our case, the baseline system to be compared against. The cut-off threshold on PD standard deviation for a given time in the video decides whether or not it is included in a summary segment. We observe that $k = 1$ already gives good selectivity to our method and at higher values like $k = 2, 3$ we detect very few and highly arousing segments. Figure 5 plots $k$ against the fraction of video that is detected as arousing by our method for two subjects for the clip *dotrc*, the trend remains consistent with increasing number of users or different clips. For values $k < 1$ our method can label upto 60 % of the stream as being arousing and hence generates too many false positives. For tighter thresholds $k > 2$, our method labels only highly arousing segments of video. We find a $k$ value close to 1 as being a reasonable operating value for our dataset and subject pool. We obtain good agreement between standard deviation based PD events from individual users and group statistics for a given clip as seen in blue traces in Figure 4 (a) and (b) respectively.

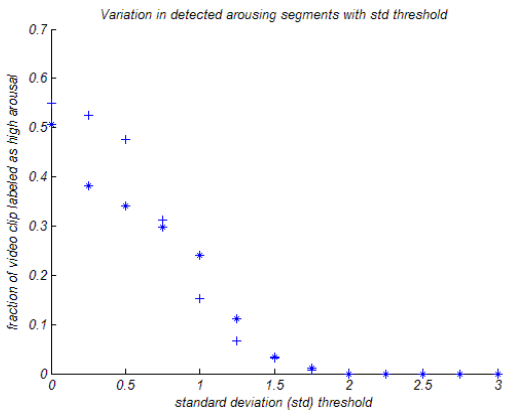
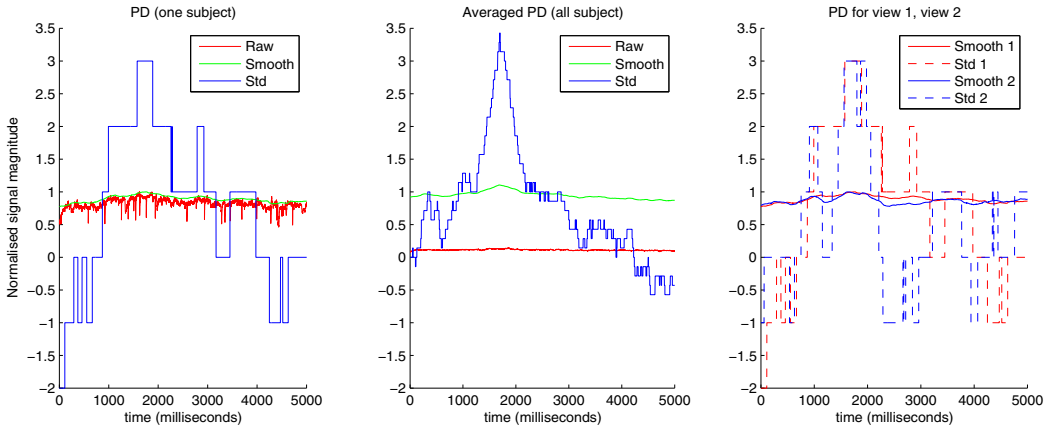



Fig. 5. Increase in $k$ threshold results in decrease in the fraction of video labeled as arousing by our method. Plots shown for PD information from two subjects watching clip *dotrc*.




(a) (b) (c)

Fig. 4. A representative result (a) PD signal processing for *jbdy* by user 14. (b) Averaged signals from all users for *jdby* (c) Smoothed signals and standard deviations corresponding to two different viewing sessions indicating arousal at similar points in the sequence.

We compute Pearson rank correlation over the PD sequences, this is an appropriate measure to compare sequences which might be scaled and shifted with respect to each other. The Pearson rank correlation gives 1 or -1 for completely correlated, and 0 for totally uncorrelated sequences. PD processing from one sequence is illustrated in Figure 4 (a). To identify temporal events of high arousal, we first smooth the signals using a moving average of $t = 4$ seconds. This value has been chosen to reflect the slower speeds of arousal variation [10]. The outlying events are then identified by binning PD values according to the standard deviations $\sigma = stdev(PD_i)$ from mean PD value as visualized in the blue trace in Figure 4 (a). Currently, we consider frame sequences with arousal values $stdev(PD_i) > k \times \sigma$ These segments are concatenated to generate a short summary video. We obtain good agreement between PD values per user, obtained from successive views of a given clip by the same user Figure 4 (c).

Story-boards are typically chosen from key frames in the video. Ideally, the subset of key frames should not only be able to represent a succinct summary of the video, but also capture the affective events in the discourse / story underlying the clip. Our PD based arousal detection helps to satisfy the latter criterion. In this paper, we rely on another behavioral cue to capture semantics in the video stream. Recent literature has shown that key objects and their interactions are responsible for a bulk of the semantics in static and dynamic social scenes [26][17]. Furthermore, eye-fixations have shown to be biased towards regions containing faces, people or other objects important to the scene context [4]. We borrow from these ideas and model the following hypothesis,

- The visual concept responsible for arousal during the given time window is very likely to be fixated upon immediately following a PD event.

This boils down to identifying the fixations immediately following a PD event. Thus we identify a single key frame corresponding to the first fixation following each PD arousal peak. These key frames corresponding are then added to the story board.

## III. Data collection and eye-tracking setup

Video data has been chosen from Youtube® to represent a variety of themes based on three parameters, extent of activity involved, spoken language in the video and whether the video has been shot indoors or outdoors. Details of the clips have been summarised in Table I. The original video clips are of 5 mins duration. The clips were normalized and re-encoded to have either a 640 pixel height or 480 pixel width, without altering the aspect ratio. Video clips were of sufficient quality for subjects to identify details such as lip movements. All clips are social scenes.

| Video clip | Theme | Location | Language | Activity level |
|---|---|---|---|---|
| **mmtat** | *Animated clip* | *Outdoor* | *English* | *medium to high* |
| **jbdy** | *Comedy scene* | *Indoor* | *Hindi* | *high* |
| **dotrc** | *Social scenes* | *Indoor and Outdoor* | *Mandarin* | *low to high* |
| **village** | *Social scenes* | *Outdoor* | *Hindi* | *medium to high* |
| **hockey** | *Sports, Hockey* | *Outdoor* | *Hindi* | *medium to high* |

TABLE I

DESCRIPTION OF THE VIDEO CLIPS CHOSEN FOR EVALUATION OF THE ONLINE FRAMEWORK AND APPLICATIONS. THE CLIPS WERE OBTAINED FROM THE PUBLIC DOMAIN AND NORMALIZED TO A 5 MINUTE DURATION. THE CLIPS WERE CHOSEN FROM AMONGST SOCIAL SCENES AND CAPTURE A DIVERSITY IN THE THEME, INDOOR AND OUTDOOR LOCATIONS, SPOKEN LANGUAGE AND EXTENT OF ACTIVITY IN EACH VIDEO CLIP.

### A. Eye-tracking setup and Subject pool

We use a desktop based ERICA ®eye tracking system. Gaze points are acquired at 30 Hz with 1 degree visual arc accuracy. Stimulus video is shown on a Dell monitor and subjects are seated at a viewing distance of 2.5-3 feet. The environment is relatively unconstrained with the presence of a few distractors in a typical laboratory working environment to have a realistic viewing setting. 20 people were recruited for the subject pool from amongst male and female graduate

and undergraduate students aged 18-32 ($\mu = 24$ years) years. There were 13 *Mandarin* speakers and 7 *Hindi* language speakers. All subjects had normal or corrected vision and were compensated with a token cash payment.

## IV. Experiment design

In the first phase, subjects are shown 9 to 10 clips of 5 minute duration over an hour long session, a short break is given between successive clips. The clips are ordered in random fashion to avoid systematic biases. The second phase of the experiment is related to the usefulness of eye-gaze data to generate summary video clips and image based story boards. Per subject eye-gaze and pupillary dilation (PD) information recorded during the first phase of the experiment is used to arousing segments of video from each clip and generate a summary of up to 1.5 minutes for each clip in Table I. Additionally, key-frames are also extracted from each clip, these represent significant events in PD information and are put together as a *story board* as illustrated in Figure 6 for the clip *jbdy*, for one subject. Subsequent to generation, video summaries as well as the story board images are then shown to the same subject and quantitative scores obtained on whether the video summary and the story board correspond with the subject's notion of significant and interesting events in the respective video clip.

## V. Results and discussion

We discuss some important results and insights from our systematic evaluation of video summary and story board generation.

### A. Video summarization

Eye gaze and PD information from each subject were analysed and two types of summary clips were generated for every participant. The first set of 4 summaries were using PD information from the subject's own viewing session for the corresponding videos, if the subject had seen the video at least once. Two clips of the second type of summary clip were generated using PD information from a different subject's viewing session. The subjects then rated the summaries for recall of interesting events in the video and how well the extracted segment of video captured such interesting events. Subjective ratings were obtained for close to a hundred such summaries indicating the following,

*1) Subjective user feedback:*

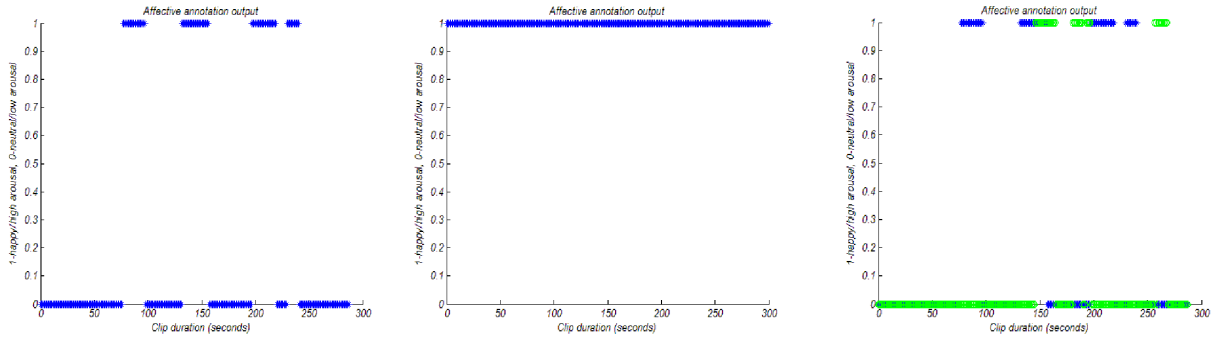
- PD based summarization captured close to 40% of interesting events in the videos shown to the user. Identifying large deviations from mean PD based arousal may not be the best strategy to extract video segments, as users indicated a significant number of video segments, that have low or moderate arousal scores.
- There is a larger subjective bias in PD based information for arousal, more analysis needs to be done to assess if this is related to the video content or to some other extraneous factor.
- There is good potential to complement PD based summarization with content based features proposed in literature [10]. This can be inferred from our observation that a good proportion of segments that were missed out by the PD based summarization have noticeable motion saliency and colour contrast variations.
- Evaluation needs to be done on longer video clips, as the current selection of clips may not be long enough to trigger sufficient number of meaningful PD events.

*2) Objective comparison with [34]:* To compare with [34], we select the clips *jbdy, dotrc, village* from our dataset. These clips have segments with positive or neutral valence. We label our videos as *high/low* arousal using PD information analysis. The time series of *high/low* arousal labels and *happy/neutral* time-series annotation by [34] is compared for agreement with the corresponding human annotated ground truth. A set of representative output annotation time series for the clip *dotrc* by applying the algorithm in [34] and our method based on PD information analysis over eye-gaze from two users (14,15) is shown in Figure 7 (b) and (c) respectively. Figure 7 (a) is the reference ground truth annotation for low/high arousal. The result from our method is illustrated for two users to show that there is good agreement between arousing content across different subjects, for a given clip. Pearson correlation coefficient values of 0.6 and above are not uncommon. We find overall that our method might miss out on some arousal events, but gives less false positives as compared to a very large number (over 70%) in [34]. Our overall analysis gives the following insights,

- We find a consistent performance improvement in our method where arousal is due to abstract semantics such as object-interaction due to actions and verbal communication.
- Unlike [34] that is agnostic to the cultural bias of a subject, PD based analysis is able to detect arousal events due to strong cultural and linguistic familiarity. We observed this in the comedy clip *jbdy* for native language speakers.
- Our method can tune personalized summaries better than the [34] as it has accesses to individual biases and preferences. This is also visible in cases where native and foreign language speakers respond differently to a video clip.
- Our PD based analysis gives noisy events corresponding to shot boundaries where there are abrupt changes including that of light intensity. The strategy of rejecting PD events near shot boundaries is able to control this to a large extent.
- The baseline method by [34] gives a very large number of false positives for the chosen clips as compared to the PD based events.

It is important to note here that the baseline algorithm [34], has its own merits in being completely automated and being able to generate summaries to cater for personalization and community preference. It is also tuned currently


Fig. 6. Panels illustrate the story board generated for a viewing session over the clip *jbdy*. The PD information is smoothed using a moving average and large deviations from the mean are isolated. Local maxima and minima are then identified as shown in the inset to the left. The key frames sampled from this process are shown in sequence (left to right, top row then second row). The frames were rated by the user to contain 6 of 10 key events. Frames are from the video *jbdy*, copyright owned by the National Film Development Center, India.




(a) (b) (c)

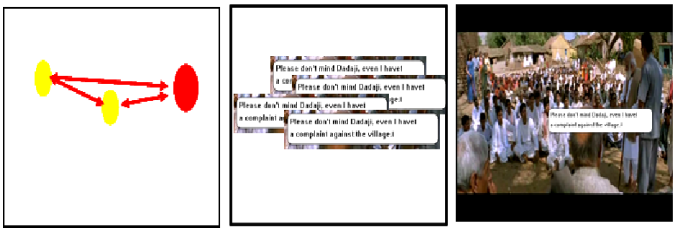
Fig. 7. Affective *high/low* arousal labels for the clip *dotrc* (a) manually annotated ground truth (b) by the algorithm in [34] and (c) from our method based on PD information analysis with $k = 1.5$, for user 14 and 15. There is good agreement between PD based arousal for the two users and manual ground truth. In contrast, [34] shows over 70% false positives.

for affective summarization of home videos. The baseline is affected significantly by the quality of a noisy and slow video segmentation step. Our method on the other hand is agnostic to underlying video segments at present. We intend to incorporate content cues along with PD information for more meaningful summaries. This would be especially valuable as the video clips become longer.

### B. Key frame based story boards

As described earlier, key frames were extracted from the video stream, at and around the local PD extremum. These were shown to user as a linear story board of images. The subject feedback for story board construction was not as encouraging as that for video summarization. We found one of the reasons to be the significant number of redundant frames. A combination of content based analysis and PD analysis is required to weed out the redundant, near-duplicate frames and is part of our future work. Our idea of mapping eye-fixations with key frames gives promising results though, ensuring that the key frame chosen has an instance of the concept related to arousal in the segment.

Another direction of current work is to convert story-boards into comic strips with dialogue overlay and expect this to improve the narrative context, akin to [14]. An example output is shown in Figure 8, (a) past(yellow disc) and current (red disc) eye-fixations (b) possible locations for dialogue overlay (c) optimal position chosen based on eye gaze based saliency. Though automatically computed saliency in [14] to determine placement of dialogues on the video key-frames, we on the other hand intend to use a novel notion of eye-movement based saliency.




(a) (b) (c)

Fig. 8. Use of eye fixations based saliency for dialogue overlay on video key-frames.

## VI. Conclusion and Future Work

We have proposed and implemented an efficient, semi-automated framework using eye-gaze to generate affective summaries.Pupillary dilation has also been introduced as a

novel and useful behavioural signal to estimate subject arousal. Arousal is an important behavioural parameter and has extensive applications in cinema, advertising, etc. Our framework and applications are also easily realizable using the new generation of hand-held devices equipped with multiple cameras and can offer a promising direction of research. We demonstrate that it is useful to invest in, and involve behavioural signals such as PD and HVA to solve a challenging video analytics problem. A deeper investigation of PD based signals establishing the correlation with existing content based features or combining with either fully automated [34] or hybrid [25] state-of-art, is another line of investigation worth pursuing. We now have a rich corpus of native and foreign language speakers watching videos in different spoken languages and this gives an opportunity to study variations of arousal patterns for subjects viewing affective video sequences.

## REFERENCES


[1] J. S. Agustin, H. Skovsgaard, J. P. Hansen, and D. W. Hansen. Low-cost gaze interaction: ready to deliver the promises. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, CHI '09, pages 4453–4458, New York, NY, USA, 2009. ACM.

[2] J. S. Agustin, H. Skovsgaard, E. Mollenbach, M. Barret, M. Tall, D. W. Hansen, and J. P. Hansen. Evaluation of a low-cost open-source gaze tracker. In *ETRA '10: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 77–80, New York, NY, USA, 2010. ACM.

[3] S. Arifin and P. Cheung. Affective level video segmentation by utilizing the pleasure-arousal-dominance information. *Multimedia, IEEE Transactions on*, 10(7):1325 –1341, nov. 2008.

[4] P. Baldi and L. Itti. Of bits and wows: A bayesian theory of surprise with applications to attention. *Neural Netw.*, 23(5):649–666, 2010.

[5] MM Bradley, L Miccoli, MA Escrig, and PJ. Lang. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, pages 602–607.

[6] M. Cerf, J. Harel, W. Einhuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Neural Information Processing Systems*. MIT Press, 2007.

[7] Paul Ekman. Basic emotions. In Tim Dalgleish and Mick Power, editors, *Handbook of Cognition and Emotion*, Sussex, UK, 1999. John Wiley and Sons.

[8] Yunlong Feng, Gene Cheung, Wai tian Tan, and Yusheng Ji. Hidden markov model for eye gaze prediction in networked video streaming. In *IEEE International Conference on Multimedia and Expo*, 2011.

[9] Andrew J. Gerber, Jonathan Posner, Daniel Gorman, Tiziano Colibazzi, Shan Yu, Zhishun Wang, Alayar Kangarlu, Hongtu Zhu, James Russell, and Bradley S. Peterson. An affective circumplex model of neural systems subserving valence, arousal and cognitive overlay during the appraisal of emotional faces. *Neurophysiologica*, (8):2129–2139.

[10] A. Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *Multimedia, IEEE Transactions on*, 7(1):143 – 154, 2005.

[11] W. Heinrich. Die aufmerksamkeit und die funktion der sinnesorgane. *Zeitschrift fur Psychologie und Physiologie der Sinnesorgane*, 9:342:388, 1896.

[12] Eckhard H. Hess and James M. Polt. Pupil size as related to interest value of visual stimuli. *Science*, 132(3423):349:350, 1960.

[13] Eckhard H. Hess and James M. Polt. Pupillometrics. pages 491–531, 1972.

[14] Richang Hong, Xiao-Tong Yuan, Mengdi Xu, Meng Wang, Shuicheng Yan, and Tat-Seng Chua. Movie2comics: a feast of multimedia artwork. In *Proceedings of the international conference on Multimedia*, MM '10, pages 611–614, New York, NY, USA, 2010. ACM.

[15] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.

[16] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[17] H. Katti, S. Ramanathan, M. S. Kankanhalli, N. Sebe, T. S. Chua, and K. R. Ramakrishnan. Making computers look the way we look: exploiting visual attention for image understanding. In *Proceedings of the international conference on Multimedia MM '10*, pages 667–670, New York, NY, USA, 2010. ACM.

[18] Duncan R. Kerr and Nicholas V. King. Systems and methods for adjusting a display based on the user's position, 12 2009.

[19] K. Kim, S. Bang, and S. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42(3):419–427, May 2004.

[20] Jeff Klingner, Barbara Tversky, and Pat Hanrahan. Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Neurophysiologica*, (3):323332.

[21] P.J. Lang, M.M. Bradley, and B.N. Cuthbert. (iaps): Affective ratings of pictures and instruction manual. technical report a-8. Technical report, University of Florida, 2008.

[22] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *J. Vis. Comun. Image Represent.*, 19(2):121–143, 2008.

[23] Pere Obrador, Rodrigo de Oliveira, and Nuria Oliver. Supporting personal photo storytelling for social albums. In *Proceedings of the international conference on Multimedia*, MM '10, pages 561–570, New York, NY, USA, 2010. ACM.

[24] Timo Partala and Veikko Surakka. Pupil size variation as an indication of affective processing. *Int. J. Hum.-Comput. Stud.*, 59(1-2):185–198, 2003.

[25] Wei-Ting Peng, Chia-Han Chang, Wei-Ta Chu, Wei-Jia Huang, Chien-Nan Chou, Wen-Yan Chang, and Yi-Ping Hung. A real-time user interest meter and its applications in home video summarizing. In *ICME*, pages 849–854, 2010.

[26] S. Ramanathan, H. Katti, R. Huang, T. S. Chua, and M. S. Kankanhalli. Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. In *ACM International conference on Multimedia 2009*, pages 729–732, 2009.

[27] Subramanian Ramanathan, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua. An eye fixation database for saliency detection in images. pages 30–43, 2010.

[28] J. A. Russel. A circumplex model of affect. *Journal of Personality and Social Psychology*, pages 1161–1178, 1980.

[29] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. Gaze-based interaction for semi-automatic photo cropping. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, pages 771–780, New York, NY, USA, 2006. ACM.

[30] Anthony Santella and Doug DeCarlo. Robust clustering of eye movement recordings for quantification of visual interest. In *Proceedings of the 2004 symposium on Eye tracking research & applications*, ETRA '04, pages 27–34, New York, NY, USA, 2004. ACM.

[31] A. Shamir and O. Sorkine. Visual media retargeting. In *SIGGRAPH ASIA '09: ACM SIGGRAPH ASIA 2009 Courses*, pages 1–13, New York, NY, USA, 2009. ACM.

[32] M. Soleymani, J.J.M. Kierkels, G. Chanel, and T. Pun. A bayesian framework for video affective representation. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1 –7, sept. 2009.

[33] Kai Sun and Junqing Yu. Video affective content representation and recognition using video affective tree and hidden markov models. In *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, ACII '07, pages 594–605, Berlin, Heidelberg, 2007. Springer-Verlag.

[34] Xiaohong Xiang and Mohan S. Kankanhalli. Affect-based adaptive presentation of home videos. In *Proceedings of the international conference on Multimedia*, MM '11, 2011.

[35] Shiliang Zhang, Qi Tian, Qingming Huang, Wen Gao, and Shipeng Li. Utilizing affective analysis for efficient movie browsing. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 1853 –1856, nov. 2009.