

HOTFLASHES: THUMBNAILING VIDEOS OF SOCIAL GATHERINGS BY DETECTING CAMERA FLASH ILLUMINATED FRAMES

Shiva Sundaram, Vladan Velisavljevic and Yujie Qin

Deutsche Telekom Laboratories, Ernst-Reuter-Platz-7, 10587 Berlin, Germany.
{shiva.sundaram, vladan.velisavljevic}@telekom.de

ABSTRACT

Automatic annotation of video clips is desired to efficiently thumbnail user generated content available in the Internet. Automatic techniques typically focus on a selected set of inherent video features (such as scene-cut or shot boundaries) that are deemed to be salient. Along this line, the presence of camera flash light illumination is a feature of interest. This event is usually triggered manually (by a photographer) within a scene and the selected frame(s) they occur in are deemed to be interesting in the recording: for instance, the appearance of a celebrity in a party. In this paper we present a method to detect video frames that contain flashes originating from still cameras in user generated video clips. We focus on designing features for flash illumination detection using various measures of luminance change within video sequences. Using the proposed method, we obtain detection performance of approximately 89% which is 8% absolute improvement over the baseline method that uses only change in average illumination. We also illustrate a case where flashes are automatically detected in video clips of social gatherings that can be used for thumbnailing and browsing.

Index Terms— Flash detection, feature extraction, content-based video analysis, video summarization, video thumbnailing, video ranking, user generated content.

1. INTRODUCTION

Web 2.0 technologies and availability of inexpensive high-performance cameras have contributed to a remarkable increase in user generated content in the Internet. Audio, video, and pictures of everyday events can be instantaneously captured, uploaded, documented and shared with millions of other users. Consequently, this has created a strong need for automatic annotation of multimedia for efficient indexing and easy, on-demand retrieval and browsing.

Video sequences contain variety of features that are exploited to automatically find meaningful segments. These include scene-changes or shot-boundaries, transitions and fades and even approaches using object segmentation and tracking [1, 2]. Many of these features are included in the post processing stage of video production and generally indicate a drastic change in content. Other approaches include using bottom-up or top-down attention models to derive a saliency score for the frames and extract the important frames for summarization [3]. In this work we focus on detecting a more subtle but prominent and commonly occurring feature: camera-flash illuminated frames in videos of social gatherings. The application we are targeting is *video thumbnailing*. Frames containing flash illumination are automatically selected from a video clip. Interesting segments within the video clips are also highlighted by calculating short-term averaging of high flash-activity frames. The selected frames and highlighted regions are then presented to the user in a graphical user interface (GUI) so as to enable her/him to

quickly browse the video.

Our motivation to detect camera-flash illumination in videos is based on the general observation that many social gatherings such as press conferences, celebrity red-carpet interviews, and sports etc., are captured using a combination of video and still-photography together by multiple persons. For example, while press conferences are captured in video for television programming, reporters and other members of the audience also use still-cameras to document certain instances of the scene. A flash illumination generated from still cameras indicates the exact instant when someone took a photograph. The abrupt illumination change is also captured in the video recording the same scene and therefore its appearance in the video sequence can be taken as a form of annotation (by the photographer) of its clip. Accurate camera flash illumination detection can therefore serve as a means to track salient frames of videos of social gatherings. Furthermore, an aggregate of these *annotation* events can be used as a measure of the importance of a segment; empirically a large number of flashes in a particular time of a gathering is an indicator of an interesting event happening (e.g. a celebrity stepping out of a limousine) and also an indirect indicator of the number of people (photographers) paying attention to it. Considering these aspects, we propose to use automatic flash detection as a means for quick thumbnailing, browsing and indexing of video clips.

To this end, in this paper we present our results on features designed for flash illumination detection in video and illustrate an example application in automatic video thumbnailing. The videos used in this work are unconstrained recordings of social gatherings publicly available at www.youtube.com.

This paper is organized as follows. In the next section, we related approaches to flash detection are discussed and the main challenges are also highlighted. In section 3 the details of our feature extraction procedure is presented. Then in section 4 we describe the database used to test our approach, wherein the experimental procedure is also outlined. Finally, the results of our experiments and the conclusions are presented in Section 5 and Section 6 respectively.

2. RELATED WORK

While there are other problems related to flash detection, in this work, we are particularly interested in scenes *illuminated by flash light* rather than detecting and segmenting flash light sources in a clip [4]. Flash affects the luminance and chrominance characteristics of the scene abruptly and locally only to a few adjacent frames across the video sequence. Existing flash detection methods commonly exploit a differential measurement of the luminance and chrominance characteristics across the temporal dimension. Yeo *et al.* [5] analyse the difference between the mean of the two adjacent frames and search for two consecutive sharp peaks, which indicate the flash light event. Heng and Ngan [6] show that the first-order difference in the feature characteristics between two adjacent frames does not satis-

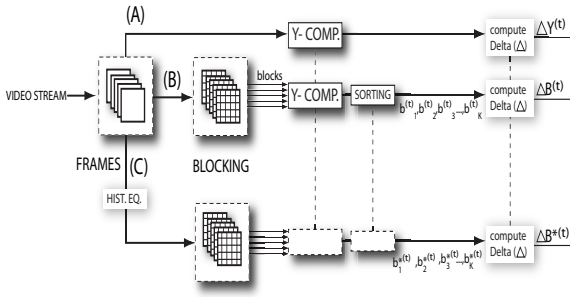


Fig. 1. Feature extraction for detecting flash illuminated scenes in video clips.

factorily discriminate the flash light from the other similar events. Hence, they extend the analysis to a series of several consecutive frames assuming the statistical characteristics of the frame features return to the previous state after the short time of flash appearance. Higher order correlation was also exploited by Sugano *et al.* [7], whereas adaptive thresholding based on intensity histogram differences was used by Zhang *et al.* [8].

A more robust method by Truong *et al.* [9] allows for detection of consecutive flash light events as long as they start with a strong luminance increase, continue with a constant luminance interval and end with the corresponding luminance decrease. In [10], a feature-based method was proposed, where the object contours in consecutive frames are matched and the intensity component is filtered across the frames. Finally, another histogram-based method for detection of both smooth and sharp illumination transitions was proposed in [4], where the difference of histograms of the consecutive frames is analyzed and classified into several classes.

One of the drawbacks of existing methods discussed so far is that they rely on the fact that the flash affects the illumination of the whole frame. As we will show in our experiments, this is not a robust approach for user generated unconstrained video content which covers a variety of scenes in indoor and/or outdoor environments and different lighting conditions and unplanned camera angles. Additionally, the location of the flash light source and the subject in a scene can vary significantly leading to cases where minor noticeable illumination changes occur (for example, on a subject's face). In contrast to the existing methods, in our work, the feature extraction procedure is based on partitioning the video frames into equal-sized blocks and subsequently using a the transition in luminance component in each block. We also combine this with histogram equalization and sorting blocks by luminance values (spatial invariance) to make detection robust for subtle illumination changes and scene changes.

3. FEATURE EXTRACTION FOR FLASH DETECTION

The feature extraction procedure adopted in this work is illustrated in Figure 1. First, a given video stream is decoded and converted an array to frames. Subsequently, the frames are processed in three separate ways: (A) The average luminance component (Y) of an entire frame is calculated. (B) Each frame is partitioned into blocks of size $N \times N$, then the average luminance component of each block (Y_{block}) is calculated and the blocks are sorted in decreasing order of Y_{block} ; sorting enables invariance to location of the illumination

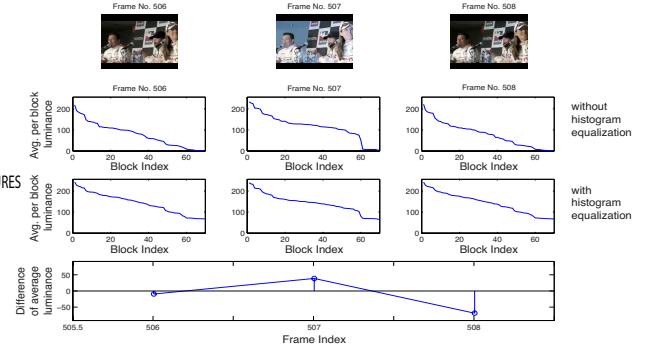


Fig. 2. Illustrations of the features obtained after blocking and sorting.

change within a frame. (C) Histogram equalization is applied to the video frames and then the blocking and sorting procedure from (B) is repeated, as illustrated in Figure 1. Histogram equalization in (C) enables adaptation and invariance to different lighting conditions across the video sequences.

Flash detection relies on abrupt change in illumination rather than the absolute values. Therefore, after sorting, in (B) and (C) the change in illumination between blocks is calculated as follows. Let $b_k^{(t)}$ and $b_k^{*(t)}$ be the k^{th} sorted average luminance value before and after histogram equalization respectively for the t^{th} frame, where $1 \leq t \leq T$ and $1 \leq k \leq K$ blocks. Then,

$$\Delta B^{(t)} = \sum_{k=1}^{k=K} (b_k^{(t)} - b_k^{(t-1)})^2, \quad (1)$$

$$\Delta B^{*(t)} = \sum_{k=1}^{k=K} (b_k^{*(t)} - b_k^{*(t-1)})^2.$$

In addition to the block-wise delta computation in (1), in the feature extraction step, we also account for the simple case where the flash saturates the luminance of the whole image frame. For this purpose we also compute the delta value for Y in (A).

Figure 2 illustrates the resulting sorted average luminance values per block for three consecutive frames. The second row illustrates the $b_k^{(t)}$ values and the third row is the $b_k^{*(t)}$ values for a given frame. From the figure it can be seen that the resulting curves change noticeably for the center frame (flash-illuminated) in comparison to the neighboring frames.

Subsequently, the delta values are augmented to obtain a 3-tuple prototype. The dimensions of this prototype are however correlated for a given video sequence. Therefore, at the end of the feature extraction procedure, the prototypes are mapped onto their principal components using principal component analysis (PCA). The features obtained are then used with a classifier to determine if a given video frame is flash illuminated or not. We are mainly interested in determining the effectiveness of the individual features for the detection task. The dataset and the experimental procedure is described next.

4. EXPERIMENTS

To estimate the performance of flash-detection implemented in our system Thirty one publicly available video clips were downloaded

Block Size	Without PCA				With PCA			
	Feature	%P	%R	%F	PCA-dims.	%P	%R	%F
8 × 8	ΔY (baseline)	73.63	92.30	81.91	dim. 01	72.41	90.03	80.26
	ΔB	76.17	89.51	82.30		-	-	-
	ΔB^*	74.49	89.81	81.44		-	-	-
	$\Delta Y, \Delta B$	83.40	92.35	87.64	dim. 01+02	80.88	92.77	86.42
	$\Delta Y, \Delta B^*$	78.58	92.93	85.15		-	-	-
	$\Delta B, \Delta B^*$	76.64	91.87	83.57		-	-	-
	$\Delta Y, \Delta B, \Delta B^*$	84.03	92.28	87.96	dim. 01+02+03	84.71	92.80	88.61
16 × 16	ΔY (baseline)	73.43	92.33	81.80	dim. 01	72.29	89.71	80.06
	ΔB	76.20	89.41	82.27		-	-	-
	ΔB^*	74.90	89.77	81.67		-	-	-
	$\Delta Y, \Delta B$	83.24	92.22	87.50	dim. 01+02	80.77	92.61	86.28
	$\Delta Y, \Delta B^*$	78.79	93.25	85.41		-	-	-
	$\Delta B, \Delta B^*$	76.53	91.70	83.43		-	-	-
	$\Delta Y, \Delta B, \Delta B^*$	82.65	93.08	87.55	dim. 01+02+03	82.52	92.60	87.27
32 × 32	ΔY (baseline)	73.44	92.50	81.87	dim. 01	71.84	87.24	78.79
	ΔB	75.31	89.71	81.88		-	-	-
	ΔB^*	74.56	90.25	81.66		-	-	-
	$\Delta Y, \Delta B$	83.11	92.46	87.54	dim. 01+02	80.29	93.25	86.28
	$\Delta Y, \Delta B^*$	78.39	93.43	85.25		-	-	-
	$\Delta B, \Delta B^*$	75.90	92.22	83.27		-	-	-
	$\Delta Y, \Delta B, \Delta B^*$	83.85	92.86	88.12	dim. 01+02+03	83.26	93.30	87.90

Table 1. Average Precision (P), recall (R) and F-measure (F) performance of flash light detection for different block size and for different combinations of the extracted features used for classification.

from www.youtube.com. The video clips¹ cover a variety of categories such as political press conferences, model photo shoots, live celebrity red-carpet interviews and concert events. The thirty one clips were manually labeled (frame-by-frame) to contain flash/no-flash (1/-1) by the authors using a graphical user interface tool. In all, 97611 frames were labeled of which 5448 were found to contain camera-flash illumination. The total duration of the dataset is 3439 seconds.

As explained in Section 3, to determine if a video frame is illuminated by flash, we extract three features and their combinations: ΔY , ΔB and ΔB^* . Based on the existing approaches discussed earlier in Section 2, the baseline approach consists of *only* using ΔY feature (change in luminance) for detection.

Classification experiments were performed to measure the performance of the individual features and their combinations. For all classification/detection tasks, we used a 3-layer Artificial Neural Network (ANN) with sigmoid activation function with R inputs and one output (Flash/Non-Flash). R varies from 1 to 3 depending on the dimensionality of the input actually used for classification. We estimated the performance of all combinations of the features and different block sizes. To estimate the classification performance for flash detection we followed a clip-wise leave-one-out procedure and averaged the results of the 31 leave-one-out estimates.

5. RESULTS

5.1. Flash illumination detection

The feature combinations and the respective performance estimates of average precision (% P), average recall (% R) and the F-measure (% F) are listed in Table 1. The table also illustrates the performances for three different block sizes: 8×8 , 16×16 and 32×32 . As shown, the left half of the table is the detection performance using the features directly (before PCA) and the right half is after transforming the features to PCA dimensions. In each case the ANN converged to less than 2×10^{-3} prediction error within 500 to 800 epochs. The dataset is highly skewed as it contains only about 5.5%

¹The data and annotation will be made available to researchers upon request

of flash-illuminated frames. This results in a chance-level precision and recall just below 70%. As mentioned earlier, our baseline performance of only using the ΔY (change in average luminance of the whole frame) feature is about 81.9%.

When using the delta measures directly with the classifier (without PCA), it can be seen that the performance of the individual measures is less than the various combinations and the best performance of about 88% is obtained when all the three measures (ΔY , ΔB and ΔB^*) are used together. In comparison to simply using the average frame-wise ΔY measure, the increased precision using the proposed measure indicates that all the three features are jointly important for robust detection of flash-illuminated frames. It can also be seen that the ΔB and ΔB^* individually perform better than ΔY , indicating the advantage of sorted block-wise analysis.

After PCA, the classification performance is estimated dimension-wise with “dim. 01” being the most significant dimension, and “dim. 03” the least. Again, it can be seen that as more and more PCA dimensions are included for the classification task, the performance improves. The change in block size has marginal effect on the performance. We obtained best detection performance for 8×8 block size. This marginal effect can be attributed to the dataset as it contains unconstrained video clips of different compression quality and different resolution.

5.2. Video thumbnailing

The screenshots in figure 3 illustrates our thumbnailing application using flash-illumination detection. Two example video clips are shown; a press conference on the left and a red-carpet event on the right. In each figure a *Heat Map* above the large image indicates the segments of flash detection; brighter red indicates higher average flash detected per unit time. On the right a list of thumbnail frames are extracted from the video clip. The user can click on the thumbnails to browse through the video. In the left screenshot, most of the heat map is dark except in two notable segments (as shown). In the video, this is where the celebrity begins to show strong emotions.

In the right screenshot, the celebrity is showing her dress to the public in the red-carpet event. In comparison to the previous video,

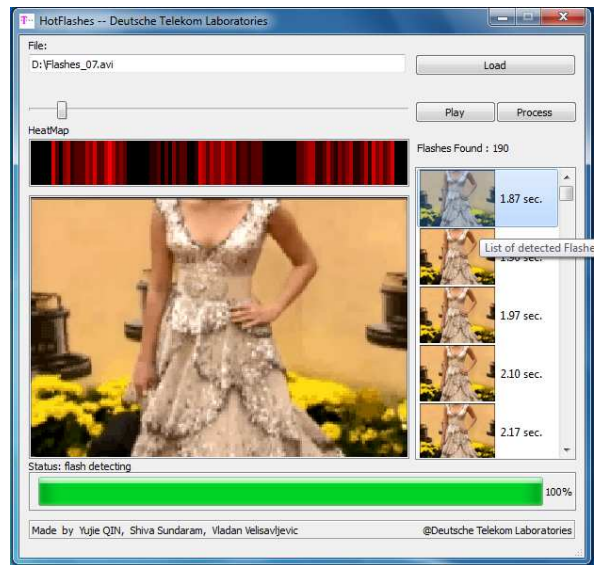
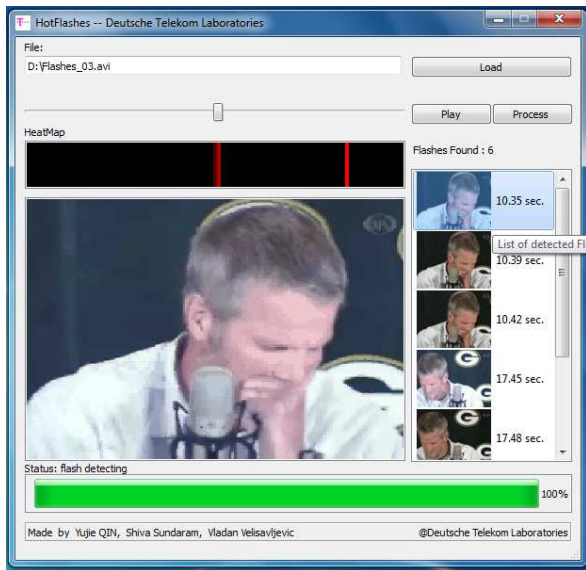


Fig. 3. Video thumbnailing using flash-detection

it can be seen that there are lot more flash detected. This is an indirect indicator of the importance/attendance of this social event compared to the press conference in the previous clip. There is also a grouping of the flash detections in to three regions in this clip, this approximately coincides with different celebrity shown in the video clip.

6. CONCLUSION

In this paper, we presented our work in detection of flash-illuminated frames in user generated unconstrained video streams. The scope of this work included defining features that leverage information in change of luminance component in video frames that are partitioned into blocks. Our main motivation to pursue the flash-illumination detection problem in video streams is supported by the need for automatic video content analysis. To this end we have illustrated a video thumbnailing application using the proposed method. Using an aggregate measure over time, we also automatically annotate the video clip with a *Heat Map* that indicates salient segments of a clip.

We believe that flash-illumination detection can be used for a variety of salient events within video clips particularly of social gathering that is. For instance, many flashes indicate the arrival of a speaker on stage, blowing out the candles in birthdays etc. In this respect, we present encouraging results as the overall performance of the proposed method is at par with other existing methods such as [4], although the set of clips used here is unconstrained and quite challenging.

One caveat in the current work is that only neighboring frames are considered and the duration of flash-illumination occurrence in consecutive video frames is not considered. Additionally, instead of fixed-size blocks, meaningful background and foreground segmentation can help detect subtle changes caused by camera flashes although it may not cause significant change in the luminance component of a given video frame. Other ideas include using attention models [3] for detecting flash illuminated frames or even regions of flash illumination within frames. These are part of our current and planned future work.

7. REFERENCES

- [1] Alan Hanjalic, "Shot-boundary detection: unraveled and resolved?," *IEEE Transaction on Circuits and Systems for Video Technology*, pp. 90 – 105, Feb. 2002.
- [2] Shu-Ching Chen, Mei-Ling Shiu, Cheng-Cui Zhang, and R. L. Kashyap, "Video scene change detection method using unsupervised segmentation and object tracking," *IEEE International Conference on Multimedia Expo (ICME)*, p. 15, Aug. 2001.
- [3] S. Marat, M. Guironnet, and D. Pellerin, "Video summarization using a visual attention model," *15th European Signal Processing Conference (EUSIPCO)*, Poznan, Poland, 2007.
- [4] X. Qian, G. Liu, and R. Su, "Effective fades and flashlight detection based on accumulating histogram difference," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 10, Oct. 2006.
- [5] B. L. Yeo and B. Liu, "Rapid scene analysis on compressed video," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 5, no. 6, Dec. 1995.
- [6] W. J. Heng and K. N. Ngan, "High accuracy flashlight scene determination for shot boundary detection," *Signal Processing: Image Communications*, vol. 18, no. 3, Mar. 2003.
- [7] M. Sugano, Y. Nakajima, H. Yanagihara, and A. Yoneyama, "A fast scene change detection on MPEGcoding parameter domain," in *IEEE Int. Conf. on Image Processing*, Chicago, IL, Oct. 1998.
- [8] D. Zhang, W. Qi, and H. J. Zhang, "A new shot boundary detection algorithm," in *Proc. PCM*, Beijing, China, Oct. 2001.
- [9] B. T. Truong and S. Venkatesh, "Determining dramatic intensification via flashing lights in movies," in *IEEE Int. Conf. Multimedia Expo*, Tokyo, Japan, Aug. 2001.
- [10] W. J. Heng and K. N. Ngan, "Integrated shot boundary detection using object-based techniques," in *IEEE Int. Conf. on Image Processing*, Kobe, Japan, Oct. 1999.