Facial Parameters and Their Influence on Subjective Impression in the Context of Keyframe Extraction from Home Video Contents

Uwe Kowalik[1], Go Irie[1,2], Yasuhiko Miyazaki[1], and Akira Kojima[1]

[1] NTT Cyber Solutions Laboratories, Nippon Telegraph and Telephone Corporation
[2] Dept. Information and Communication Engineering, University of Tokyo
{kowalik.uwe,irie.go,miyazaki.yasuhiko,kojima.akira}@lab.ntt.co.jp

Abstract. In this paper, we investigate the influence of facial parameters on the subjective impression that is created when looking at photographs containing people in the context of keyframe extraction from home video. Hypotheses about the influence of investigated parameters on the impression are experimentally validated with respect to a given viewing perspective. Based on the findings from a conducted user experiment, we propose a novel human-centric image scoring method based on weighted face parameters. As a novelty to the field of keyframe extraction, the proposed method considers facial expressions besides other parameters. We evaluate its effectiveness in terms of correlation between the image score and a ground truth user impression score. The results show that the consideration of facial expressions in the proposed method improves the correlation compared to image scores that rely on commonly used face parameters such as size and location.

Keywords: Keyframe extraction, home video contents, image ranking, facial expressions.

1 Introduction

Due to the increasing popularity of consumer video equipment the amount of user generated contents (UGC) is growing constantly in recent years. Besides taking photographs also filming at family events or during travel became a common habit for preserving valuable moments of one's life. Whereas taking a good picture of a person with a photo camera requires a decent skill and good timing, a video can easily capture 'good shots' due to its nature. Thus extracting images from video contents is an interesting complement to the task.

In recent years the problem of automatically extracting representative keyframes from video contents attracts the attention of many researchers and various approaches have been proposed to tackle this sub-problem of video abstraction. Existing keyframe extraction techniques can be categorized based on various aspects such as underlying mechanisms, size of keyframe set and representation scope [1]. From the viewpoint of analyzed base units, there exist two

S. Boll et al. (Eds.): MMM 2010, LNCS 5916, pp. 76–86, 2010.

categories: shot based extraction methods and clip based methods. Whereas shot based approaches involve always a shot detection step and thus are limited in their application to structured (i.e. edited) video contents, clip based methods can be applied to unstructured video material such as user generated contents. Despite the fact that there exists a huge variety of proposed keyframe extraction approaches, little attention has been paid to the fact that extracted keyframes should also be visually attractive to the viewer.

In this paper, we address this issue for the home video content which is one specific domain of UGC. Home video contents contain often imagery of people, relatives and friends taken at various occasions and therefore human faces are intuitively important. In addition, the human faces do attract a viewer's attention [2,3]. Provided an application context where the objective is to extract frames from videos with respect to the evoked impression, such as creating a family's photo album and sending picture e-mails, a fully automatic keyframe extraction approach will ideally select images that are considered as 'good shot of the person(s)'. The central problem is how to automatically determine such 'good shots' inside video sequences. An important point is that selected video frames should suffice a certain image quality. This can be achieved by standard techniques such as image de-noising, contrast and color enhancement [4], advanced methods for increasing the image resolution [5], and image de-blurring [6]. In contrast, we focus in this paper on intrinsic image properties that can not be easily changed without altering the image semantics.

More specifically, main contributions of this work are:

- To investigate the influence of facial parameters present in images on the viewer's impression in the context of keyframe extraction from home video contents
- To present results of a conducted subjective user experiment
- To propose an image scoring method based on a weighted combination of extracted face parameters

As a novelty to the field of keyframe extraction, the proposed image score considers facial expressions besides other parameters. We discuss different face parameter combinations and their influence on the performance of a general linear weighting model used for image score estimation. We show results that confirm the effectiveness of included facial expression parameters.

2 Related Work

We focus in this section on the conventional methods that aim on keyframe extraction from unstructured video such as UGC.

In [7] a keyframe extraction approach suitable for short video clips is introduced in the context of a video print system. Keyframes are determined by employing face detection, object tracking and audio event detection. Location of visual objects is taken into account by predefining regions of importance inside the video frame and deriving an appropriate region based frame score. Although

the parameters of visual objects (and especially faces) are rather intuitively considered for frame score calculation in the above approach, the example illustrates the awareness of visual object importance in the context of automatic keyframe extraction.

An interesting study focusing on unstructured user generated contents was recently presented in [8]. The authors conducted a psycho-visual experiment in order to derive common criteria for human keyframe selection shared amongst two user groups with different viewing perspectives i.e. first-party users (photographers) and third-party users (external observers). A rule-based keyframe extraction framework is proposed that uses inference of the camera operator's intend from motion cues derived from estimated camera and object motion patterns rather than recognizing the semantic content directly. The authors compare their method with a histogram based and uniform sampling keyframe extraction approach and showed the effectiveness of the algorithm by improved accuracy with respect to the ground truth selected by a human judge. The authors suggest that the accuracy can be further improved by inclusion of semantic information such as facial expressions and gaze at camera. In contrast to the system proposed in [8], this paper focuses on such highly subjective parameters that influence the user's keyframe choice wrt. the *attractiveness* rather than approaching the keyframe extraction problem with the goal to extract interesting images from the video.

In [9], a user-centric weighting model for image objects is proposed that defines an importance measure for images based on users' perception. The authors investigated the relationship between parameters of visual objects such as size, location and the user's perception of 'image aboutness' with respect to one given, object specific query term. It was shown that object size and object location are directly related to the image concept recognized by a user and that their proposed weighting model is efficient for image ranking in the context of concept based image retrieval.

In this paper we view the task of keyframe extraction as a conceptual image ranking problem within a given video sequence. We focus on home video contents since there is a wide range of potential applications for human-centric video indexing technology in this domain. The goal of this study is to provide some insight in how the presence of faces in images influences the viewers' impression in terms of 'a good photograph of a person'.

3 Impression Concept and Investigated Face Parameters

In this work, we define the impression evoked at viewers when looking at a photograph as based on the concept of 'a good picture of a person'. It is assumed that various face parameters contribute to the impression with respect to this conceptual viewpoint. We assess two different face parameter types with regard to their contribution:

1. Image structure related parameters
2. Emotion related parameters

Image structure related face parameter considered in this work are *number of faces*, *face coverage* and *face location*. As a novelty to the task of keyframe extraction, we consider also *emotional* face parameters. We model emotional face parameters by a prototypic facial expression class label assigned to each detected face region according to Ekman's six basic emotions [10]. In particular we focus on two prototypic facial expressions *joy* and *neutral* in this study. Joyful faces are commonly considered to be attractive to viewers. The neutral expression is important, since it reflects the 'normal' state of human faces. In the following, we present our assumptions regarding the relationship between each facial parameter and the impression evoked at viewers where the impression is quantified by means of a user provided impression score.

Face Number N_f: As suggested by related work we assume that images containing more faces are considered to create a better impression at the viewer and thus the *face number* is positively correlated with a user provided impression score.

Face Coverage S: We define face coverage (hereafter: *coverage*) as the ratio between the image area covered by faces and the overall image area as:

$$S_{image} = \frac{1}{A^{image}} \sum_i A_i^{face} \tag{1}$$

where A^{image} denotes the image area and A_i^{face} is the image area covered by i-th face. We assume that larger faces evoke a stronger impression at viewers and thus coverage has a positive correlation with a user provided impression score.

Face Location P_R: We use a region of interest (ROI) approach for describing the face location. P_R is defined with respect to three predefined ROI based on a bary center of a face's rectangular bounding box. Fig. 1 depicts the predefined regions. Our assumption is that there exist preferred ROIs that will lead to a better impression if a face lies inside such a region.

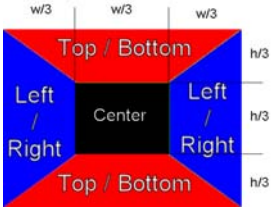

Fig. 1. Three predefined ROI (w: image width, h: image height)

We model the face location P_R by the probabilities that a face lies inside a region R as given in the formula:

$$P_R = \frac{N_f^R}{N_f} \tag{2}$$

where $R \in \{top/bottom, center, left/right\}$ and N_f^R refers to the number of faces inside the region R. We define $P_R = 0$ if $N_f = 0$.

Facial Expressions P_E: Our hypothesis is that facial expressions present in images are important for the overall impression. We assume that the presence of joyful faces will influence the impression positively whereas a present neutral facial expression will have little or no correlation with the user score. We parameterize facial expressions by their occurrence probability P_E in a frame as given in the fomula:

$$P_E = \frac{N_f^E}{N_f} \tag{3}$$

where $E \in \{neutral, joy\}$ and N_f^E equals the number of faces displaying the expression E. This general formulation allows for future extension by adding other prototypic facial expressions. We define $P_E = 0$ if $N_f = 0$.

4 Subjective User Experiment

Goal of the user experiment was to acquire ground truth data in form of a score that quantifies the subjective impression evoked at the participants when looking at the extracted images. We prepared two facial parameter sets based on the images used for the experiment. The first set was manually labeled and provides ground truth data for parameter assessment under the assumption of an ideal system. In order to be able to draw conclusions about the impression-parameter relationship under practical conditions, the second parameter set was generated during the fully automatic keyframe extraction process used for preparing the test images. In the following subsections, we first describe the data preparation, and next provide a description of the experimental conditions.

4.1 Selection of Video Clips

Video clips of three different categories were selected from private home video collections and a video sharing website [12]. We consider creating photo albums from home video as one of the main applications for keyframe extraction techniques and choose therefore two popular categories which roughly relate to the 'home photo' taxonomy proposed in [13], i.e. *travel* and *family Events*. Investigating the thumbnail previews in the family category of [12], we found that nearly 80% of the uploaded video clips are about children and decided therefore to add a special *kids* category to our test set. We selected two typical clips from each of three categories resulting in overall six video clips. The clip properties are listed in Table 1. The number of extracted images per clip used for the experiment is also listed. Fig. 2 illustrates the video contents by providing one example for each category.

4.2 Automatic Keyframe Extraction Approach

Fig. 3 shows the block diagram of the keyframe extraction method. In order to avoid quality degradation by motion blur, the video is first analyzed to detect

Table 1. Video Clip Properties

Category	Title	Length (min)	Num. of Frames	Resolution / FPS	Num. Extracted Frames
Travel	seaside	12:17	11000	320x240 / 15	23
	disney land	02:40	4800	320x240 / 30	7
Family Events	birthday	04:50	8600	320x240 / 30	21
	exhibition	02:40	4800	320x240 / 30	4
Kids	baby	02:00	3700	320x240 / 30	4
	stairs	00:50	1400	320x240 / 30	5

Fig. 2. Example images (left to right: travel, family event, kids)

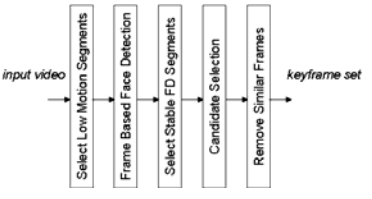

Fig. 3. Keyframe Extraction Process

strong global motion patterns by estimating a segment based motion activity, which is calculated for non-overlapping consecutive groups of N video frames by evaluating the structural similarity between neighboring frames. As for the structural description, we employ a feature vector constructed from Gabor-Jets extracted at equidistant grid positions. We employ the similarity measure introduced in [14] for calculating the similarity between adjacent video frames I_k and I_{k+1}. As for the Jet-configuration, we use five frequencies and eight orientations, the same parameters given in [14].

Video segments that contain high motion activity show usually a low similarity between adjacent frames. We calculate the average similarity for each segment of N video frames and apply a threshold th. Segments holding an average similarity $< th$ are discarded. In our implementation $N = 10$ and $th = 0.9$ lead to good blur suppression results. Only video segments with low motion activity are preselected for further processing and subjected to frame-wise face detection based on the approach introduced in [15]. The number of candidate frames is further reduced by removing video frames with unstable face detection result. We apply a sliding window function of length $K(K = 3)$ and remove frames where

the face count in range divided by K is less than 1.0. A redundancy removal step finally removes visually similar frames in sequential order by calculating the inter-frame similarity of adjacent frames as described above and applying a threshold ($th_2 = 0.7$). Frames with a similarity higher than th_2 are removed and the remaining set of keyframes was used for the subjective experiment. Overall 64 images were automatically extracted from the six video clips. Based on the approach given in [11], we have implemented an automatic facial expression detection module for the detection of neutral and joyful facial expressions. It utilizes a neural network for classifying facial expressions based on Gabor-Wavelet feature vectors extracted from detected face image regions. Facial expression detection was applied on the extracted images in order to create the automatically extracted face parameter set mentioned above.

4.3 Experimental Condition

In our experiment, we asked subjects to watch the extracted video frames and provide their opinion with respect to a given question about the images. The experiment was performed at an university amongst students and staff members. Participants were unrelated to the people included in the imagery thus the judgment was given from a 'third-party' viewpoint. The overall number of participants was 22 (19 males, 3 females) in the age between 21 to 49 years. Images were displayed in the center of the computer screen in original size and in random order. Participants were asked to give their feedback with respect to the question

"Do you think this picture is a 'good photograph' of the person?"

The feedback was given directly after watching each single image by selecting a score on a 7-level ordinal scale ranging from -3 for *'not at all'* to +3 for *'I fully agree'*.

5 Experimental Results and Discussion

In this section, we first investigate the relationship between each single face parameter and the subjective user score and draw conclusions regarding the validity of the assumptions made in section 3. This investigation is performed under the assumption of an ideal system, i.e. we use the ground truth face data. Based on the results, we introduce next a general weighting model which uses the validated face parameters for calculating an image score that considers the subjective impression of viewers and is useful for human-centric image ranking with respect to the viewing perspective previously defined. We show the performance of different score predictor combinations in terms of rank-correlation between the predicted image score and the impression score acquired during the user experiment from section 4. Finally we select the best parameter combination and compare the performance of the weighting model for ideally labeled face parameters and the practical case where face parameters are automatically estimated.

5.1 Single Feature Correlation

We use the Spearman rank-correlation measure because the user score acquired during the experiment is ordinal-scaled. In order to remove a possible bias that may have been introduced due to individual differences of the participants, we calculate the normalized average score. Since the original user score is an ordinal scaled value, we decided to calculate the correlation for the median score as well. Table 2 shows the correlation between user score and ground truth face parameters for both score types. In addition the one-side p-value for each feature is calculated for assessing the statistical significance of the correlation value. As can be seen the correlation does not differ much between normalized average score and median score. Thus using either score type is valid. The following discussion refers to the ground truth median user score (hereafter 'user score'). As for the investigated image structure related face parameters only the coverage feature shows statistically significant correlation. The other features, i.e. number of faces and face location show no or little correlation, but more important the correlation is statistically not significant. Thus our hypotheses regarding these face parameters is not validated. A similar result was presented by the authors in [9] and we conclude that our assumption is validated that bigger faces are not only more important to the user, but also contribute to a better overall impression of a photograph, whereas face location and number of faces are not valid features for deriving an image score which relates to the viewers' impression of 'a good photograph of a person'.

Table 2. Correlation of Face Parameters and Median User Score (valid features bold)

Category	Feature	Norm. Avg. Score		Median Score	
		rho	p-one	rho	p-one
Structure	Num Faces	-0.1283	0.156	-0.0876	0.246
	Center	-0.0061	0.480	0.0785	0.269
	Top/Bottom	0.1029	0.211	0.0164	0.449
	Left/Right	-0.1140	0.186	-0.0779	0.269
	Coverage	**0.2410**	**0.027**	**0.2270**	**0.035**
Emotion	**Neutral**	**-0.4183**	**<0.001**	**-0.4497**	**<0.001**
	Joy	**0.4183**	**<0.001**	**0.4497**	**<0.001**

Facial expressions seem to have a quite strong influence on the viewers' opinion. Our assumption that a joyful facial expression contributes positively to the impression was confirmed by the statistically significant positive correlation. Even more interesting we found that neutral facial expressions create a negative impression when looking at a photograph of a person. We conclude that facial expressions are an important factor of the impression created when looking at human photographs. We explain the relatively small correlation values by the influence of other image properties not discussed in this paper. Participants stated after the experiment that they considered also the gaze direction and overall image quality for giving their score. We will investigate these parameters in future work.

5.2 Linear Model for Image Scoring Based on Face Parameter Weighting

We propose a novel image scoring method that takes into account the subjective impression of a viewer evoked by the presence of faces in images by utilizing our findings from previous experiment. The score is calculated as a linear combination of the validated parameters introduced in section 3 based on the general and extendable linear weighting scheme given in the following equation:

$$S(I) = \sum_{i=1}^{N} w_i X_i = w_{coverage} S_{image} + \sum_{E \in \{neutral, joy\}}^{K} w_E P_E \qquad (4)$$

where $S(I)$ refers to the image score and $w_i X_i$ are the weighted face parameters. N is the number of face parameters and K refers to the number of facial expressions respectively. We estimate the weights w_i based on the user score by standard multiple linear regression.

Fig. 4 shows the correlation and the related p-values calculated from ground truth face parameters. Results for single feature predictors and the 99% confidence limit are also shown for convenience. The correlation value for the facial expression features is as twice as high as the result for the face coverage feature. Moreover, the correlation between user score and predicted image score is statistically highly significant which confirms our conclusions from 5.1 and we state that facial expressions are an important feature for calculating an impression related image score.

Combining *joy* and *neutral* expression predictors does not improve the correlation. We explain this by the fact that we use only two expression detectors and therefore these two features are 100% negatively correlated. Thus we gain no additional information by including both predictors. This will change, when more facial expressions are added. A calculated cross-correlation between *joy* and *neutral* expression features of $r_{jn} = -1.0$ justifies this statement. A performance improvement is achieved by combining *coverage* and facial expression based predictors. The cross-correlation between *coverage* and either of the facial expression parameters was calculated as $r_{cj} = -0.012$ and $r_{cn} = 0.012$ respectively. Thus we can expect an improvement by combining these predictors in our model.

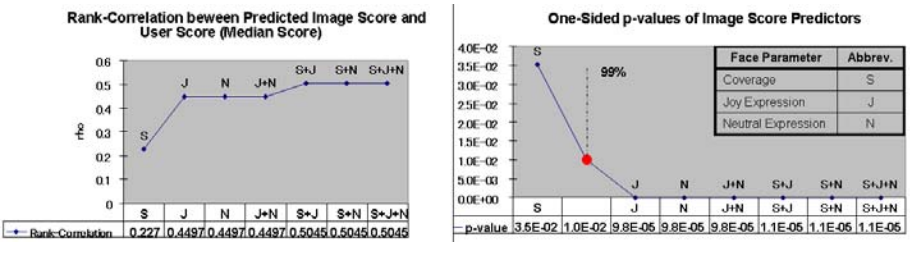

Fig. 4. Rank-Correlation and p-Values for Parameter Combinations

Based on these findings for the ground truth data, we conclude that the combination of the structural image features of *coverage* and the emotional image feature facial expressions leads to the best score prediction result. Therefore, we use this combination in our weighting model for assessing the practical case where all facial parameters are estimated fully automatically. The comparison result is given below:

- Ground Truth : rank-correlation rho=0.5, p=0.00001 (one-sided)
- Automatic : rank-correlation rho=0.34, p=0.003 (one-sided)

The correlation between the image score estimated from automatically estimated face parameters drops by 0.16 compared to the result calculated from ground truth. Analyzing the reason for this, we calculated the cross-correlation r_{xy} (Pearson) between ground-truth labeled and automatically detected face parameters. The results are:

- *coverage* $r_{xy} = 0.75$
- *joy/ neutral* $r_{xy} = 0.65$.

We conclude that the facial expression detection result contributes most to the performance degradation in our automatic system. We will address this issue in our future work.

6 Conclusion and Future Work

We investigated the influence of facial parameters on the subjective impression evoked at viewers when looking at photographs containing people from a 'third-party' viewing perspective. In the present study we focused on images extracted from home video contents in the application context of automatic keyframe extraction. Image structure related parameters such as *face number*, *face coverage* and *face location* were considered. We also investigated the contribution of facial expressions to the viewer's impression. As the results of conducted user experiments, we validated our hypotheses regarding the positive influence of coverage and joyful facial expression at the impression with respect to the predefined viewing concept of 'a good picture of a person'. Moreover, we found that the presence of neutral facial expression influences the impression negatively. The hypotheses about the existence of preferred locations for faces as well as the contribution of the face number were not confirmed, thus we conclude that these parameters do not contribute to the evoked impression. Then, we proposed an extendible linear weighting model that exploits present facial properties for calculating an image score that is correlated to the viewers' impression, and validated its effectiveness for image retrieval tasks.

An open issue to be addressed in the future is the combination of our approach with traditional keyframe extraction methods in order to determine the degree of improvement that can be achieved when emotional cues are taken into account. The relationship with persons depicted in the imagery also influences the keyframe selection and therefore we would like to address the 'first-party'

viewing perspective in the future in order to determine how a personal relationship could be possible modeled by using face parameters. Furthermore we are interested in the validation of our model for other facial expressions and image parameters in order to extend our proposed weighting scheme for image score calculation by including these parameters which we expect to increase the correlation between the predicted image score and the subjective impression with respect to the viewing perspective given in this paper.

References

1. Truong, B.T., Venkatesh, S.: Video abstraction: A systematic review and classification. ACM TOMCCAP 3, 1 (2007)
2. Gale, A.: Human response to visual stimuli. In: Hendee, W., Wells, P. (eds.) The Perception of Visual Information, pp. 127–147. Springer, Heidelberg (1997)
3. Senders, J.: Distribution of attention in static and dynamic scenes. SPIE 3016, 186–194 (1997)
4. Russ, J.C.: The Image Processing Handbook. CRC Press, Boca Raton (2006)
5. Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: a technical overview. IEEE Signal Processing Magazine 20(3), 21–36 (2003)
6. Fergus, R., Singh, B., Hertzmann, A., Roweis, S.T., Freeman, W.T.: Removing camera shake from a single photograph. In: ACM SIGGRAPH 2006, pp. 787–794 (2006)
7. Zhang, T.: Intelligent Keyframe Extraction for Video Printing. In: Proc. of SPIE Conference on Internet Multimedia Management Systems V, vol. 5601, pp. 25–35 (2004)
8. Luo, J., Papin, C., Costello, K.: Towards Extracting Semantically Meaningful Key Frames From Personal Video Clips: From Humans to Computers. IEEE Trans. Circuits Syst. Video Techn. 19(2), 289–301 (2009)
9. Martinet, J., Satoh, S., Chiaramella, Y., Mulhem, P.: Media objects for user-centered similarity matching. Multimedia Tools Appl. 39(2), 263–291 (2008)
10. Ekman, P., Keltner, D.: Universal facial expressions of emotion. In: Segerstrale, U., Molnar, P. (eds.) Nonverbal Communication, pp. 27–46. LEA, Mahwah (1997)
11. Kowalik, U., Hidaka, K., Irie, G., Kojima, A.: Creating joyful digests by exploiting smile/laughter facial expressions present in video. In: International Workshop on Advanced Image Technology (2009)
12. ClipLife, `http://cliplife.goo.ne.jp/`
13. Lim, J.H., Tian, Q., Mulhem, P.: Home photo content modeling for personalized event-based retrieval. IEEE Multimedia 10(4), 28–37 (2003)
14. Wiskott, L., Fellous, J.-M., Kruger, N., von der Malsburg, C.: Face Recognition by Elastic Bunch Graph Matching. IEEE Trans. PAMI 19(7), 775–779 (1997)
15. Ando, S., Suzuki, A., Takahashi, Y., Yasuno, T.: A Fast Object Detection and Recognition Algorithm Based on Joint Probabilistic ISC. In: MIRU 2007 (2007) (in Japanese)