# AUTOMATIC FREEZE FRAME DETECTION FOR VIDEO PRESERVATION

*Peter Schallauer, Hannes Fassold, Martin Winter, Werner Bailer*

JOANNEUM RESEARCH, Institute of Information Systems & Information Management, Graz, Austria

## ABSTRACT

A significant amount of work in film and video preservation is dedicated to quality assessment of the content to be archived or re-used out of the archive. This paper proposes automatic content analysis algorithms which reduce manual inspection time in software based preservation environments. We list the requirements for such algorithms and tools and show exemplarily on a freeze frame impairment detector how analysis algorithms need to be designed for meeting the requirements. The evaluation has shown that robustness against other impairments (noise and flickering) is an essential part of the algorithm. Successful detection of freeze frame impairments with a minimum length of three frames is achieved. The consideration of human perception is important to achieve low false detection rates. Analysis results are represented in a MPEG-7 standard compliant way. The proposed defect summary visualization tool enables efficient human exploration of visually impaired content.

***Index Terms—*** *Preservation, freeze frame impairment, metadata, summarization.*

## 1. INTRODUCTION

Automatic quality analysis of audiovisual content is an important tool in several steps of the media production, delivery and archiving process. Broadcasters are checking audio and video quality as part of the ingest process, after editing, after encoding and before play-out for terrestrial, satellite and cable broadcast or for delivery to internet and video-on-demand services. Archives are checking for content integrity at archive ingest and delivery. Content providers are checking post production content for correct encoding and conformance to the required quality and format standard before dispatching to the broadcasters or other service providers. These use cases have in common that mainly technical properties of the material are checked, e.g. stream compliance, GOP structure, playtime, aspect ratio, resolution or MXF compliance. Additionally only some content properties are checked, e.g. blocking, luma/chroma violation or noise level. In this work we focus on content based quality analysis.

Within the digital video and film preservation application domain the results of content based quality analysis aim at improving efficiency of various archive related processes. In the archive ingest process it is of interest whether content has minimum quality to be archived (e.g. check for image instability, out of focus, or freeze frames). For archive migration it is of interest whether the content quality has not been degraded due to transcoding from the legacy to the new encoding, e.g. due to blocking. Quality analysis can be used to detect the best quality copy in the case that several copies of the same content are available within the archive.

In archive exploitation it is of interest whether content has high enough quality for a certain intended usage (e.g. resolution, image stability, or noise level).

Content based visual quality analysis for the purpose of improving efficiency of preservation processes has not been specifically researched so far. In section 2 we outline requirements on algorithms and tools developed for that field of application. In section 3 we propose a freeze frame impairment detection algorithm. Section 4 evaluates how the requirements are fulfilled by this algorithm. In section 5 we propose an impairment description metadata format based on the MPEG-7 standard allowing exchange and usage of analysis results in a flexible way and present visualization and summarization techniques for efficient human interaction with quality analysis results. Section 6 provides conclusions and future work.

## 2. REQUIREMENTS ON ALGORITHMS

In previous work we have analyzed the requirements for defect and quality analysis as an appraisal step in the restoration workflow [1], viz. providing an abstracted description of the impairment and a flexible and extensible analysis framework that integrates with restoration tools. The requirements for preservation differ mainly due to the throughput requirements and the lack of user intervention. Due to the amount of content in audiovisual archives the average throughput of a system should be at least real-time and should be scalable by adding more processing nodes. Human intervention needs to be reduced to the absolute minimum, e.g. selecting an analysis profile for a batch of content but no interaction during the analysis.

The interdependencies between different impairments pose a special challenge for implementing an automatic system with acceptable detection performance. For the development of the individual impairment detection algorithms this requires robustness against influences from other impairments present in part of the material. This also requires adaptation of the algorithm to other impairments in a certain segment of the content.

For each of the algorithms the aim is to report analysis results correlating with human perception of the impairment. Thus the consideration of content properties and other defects is important as they might significantly influence the perception of the impairment strength. This means that the final assessment of each of the impairments cannot be done independently but needs to be done on higher level, taking the interdependencies between the different impairments into account.

Quality analysis for preservation shares some of the requirements with quality analysis for restoration, but adds some hard requirements in terms of throughput and correlation with human perception.

## 3. FREEZE FRAME DETECTION ALGORITHM

To fulfill the requirements stated above, the proposed algorithm for detection of freeze frame defects is composed of several components. A crucial point is the development of fast (due to real-time requirements) and robust measures for estimating the amount of content change between two images. For this purpose, we propose two different activity measures: The *visual activity* measure is designed to be very sensitive to slight content changes. At the same time it needs to be robust against noise and slight brightness changes. These impairments appear in almost all kind of material and are not perceived as content change by a human observer. As a complementary measure we use the *block motion activity* described in section 3.2. This measure is calculated from the amount of block motion between two images. Block motion activity is useful for sequences with low-contrast, slow-moving content, where the visual activity measure alone reports a very low activity value (e.g. during a slow pan showing a cloudy sky).
Using these activity measures, we first identify temporal segments with sufficiently small activity as potential freeze frame segments. Then the temporal neighborhood of the potential freeze frame segment is analyzed and a final decision is made whether to keep a potential freeze frame segment or to discard it by taking knowledge about human perception into account.

### 3.1 Visual activity

In the following we give a more detailed description of visual activity $A(I_1,I_2)$ calculation between two images. As a first step both images are converted to 8 bit gray-value images and $D = |I_1 - I_2|$ is calculated. In order to reduce noise, it is blurred with a 7x7 box filter, resulting in $D_{blurred}$ . Using its histogram $H(d)$ of gray-values, the mean $\mu$ and standard deviation $\sigma$ of $D_{blurred}$ are calculated. As a safeguard, the mean is clipped to the range [0.5, 5] and the standard deviation to [0.2, 2]. Afterwards, a noise threshold $T$ = round ($\mu + 3\sigma$) is calculated, which separates absolute differences corresponding to noise and brightness variations from 'outliers' which are most likely corresponding to image content change. Now we calculate the number of outlier pixels as $S_{out} = \sum_{d=T}^{255} H(d)$ and the average absolute difference of the outlier pixel, shifted by $T$, as $\mu_{out} = \frac{1}{S_{out}} \sum_{d=T}^{255} H(d) \cdot (d - T)$. Both values are normalized to $S'_{out}$ and $\mu'_{out}$. The final activity value is calculated as the weighted geometric mean $A = 100 \cdot (S'_{out})^{0.6} \cdot (\mu'_{out})^{0.4}$. Note the relative overweighting of the measure $S'_{out}$ as it is less sensitive to the image contrast, which can vary across the material.

### 3.2 Block motion activity

For calculating the block motion activity $B(I_1,I_2)$ in uncompressed video the images are first divided into blocks of approximately 100x100 pixels. We calculate for each block its horizontal and vertical integral projection (which are one-dimensional vectors) and use them for matching. Note that the usage of integral projections is not only advantageous in terms of computation time; it is also more robust with respect to image noise [2]. The matching itself is done separately for each direction, by minimizing the SSD score of the integral projection vectors. We simply calculate the SSD score for each integer translation value in the range -10 to 10 pixel and take the minimum. Having determined the translation for each block, we calculate the block motion activity $B(I_1,I_2)$ as the average block translation magnitude.

### 3.3 Detection of potential freeze frame segments

For detecting *potential* freeze frame segments (FFS) throughout the video, the minimum length $l_{min}$ of them needs to be defined by the user (see section 4.3). We also define thresholds $T_A$ = 0.6 and $T_B$ = 0.2 for the maximum visual activity and block motion activity for a segment to be identified as potential FFS. $T_A$ and $T_B$ are normalized measures with respect to image resolution and are constant for different type and activity of content. The start of a freeze frame segment is detected by checking for the last $l_{min}$ frames whether all the visual and block motion activity between these frames is smaller than the defined activity thresholds. In order to do as few activity calculations as possible, first the activity between the most distant frames is calculated (which is likely to be the highest within these $l_{min}$ frames). If a FFS start was encountered, subsequent frames are added to the FFS as long as the activity between the added frames and the FFS start is sufficiently small.

### 3.4 Temporal analysis

After having identified the potential freeze frame segments (FFS), we analyze them temporally. For each pair of temporally neighboring FFS (e.g. with a gap length < 5 frames), we calculate the *inner* activity for all frames within the gap zone and also the *outer* activity between the temporally most distant frames of the FFS. If the inner and the outer activity are both sufficiently small, the segments are merged. This helps to reduce the splitting of a actual freeze frame impairment into multiple freeze frame segments. In contrast, temporally neighboring FFS which have a sufficiently small inner activity but significantly higher outer activity are discarded, as this indicates a scene with very slow motion or a slow dissolve. Note that neighboring FFS, which have significant inner activity, are kept as separate freeze frame segments.

### 3.5 Human perception approximation

In the final step, we take into account human perception by noticing that short static sections often appear in video content (e.g. an actor who stands still for a moment) and shall not be detected as a freeze frame impairment. To discriminate them, the visual activity between the last frame of the potential FFS and its successor is analyzed. If it is non-significant, then the potential FFS is classified as static section and is discarded. As the normal visual activity measure is too sensitive for this purpose, we have to make the visual activity measure coarser by blurring the input images with a 5x5 box filter. We use a visual activity threshold $T_A$ = 2.0 for discriminating between significant and non-significant motion.

## 4. EVALUATION

To evaluate the performance of our novel algorithm we use a set of 8 videos (about 2.3 hours in total) for which 45 freeze frame impairments as well as all other visible impairments have been manually annotated. The videos show completely different activity

characteristics (for example a lot of high speed motion in sport scenes, slow pans in weather surveillance camera transmissions) as well as various other impairments such as noise or flicker. Our performance measure follows the classical way of true/false positive/negative determination by comparing the detected freeze frame segments of the automatically processed video to the manually annotated ground-truth. From those basic values other commonly used performance measures such as precision or recall can be deduced easily. For the matching of the time segments we allow a tolerance of 4 frames in order to tolerate temporally slightly imprecise annotations or decoders. It is important to note, that due to human scene interpretation static content in videos (e.g. photos, captions, full screen logos, end titles or black frames) is not interpreted as freeze frame impairment. Therefore such static content segments are excluded from this freeze frame impairment evaluation.

### 4.1 Basic evaluation

To get a first impression about the performance of the proposed algorithm we processed the whole dataset with the algorithm described in sections 3.1 through 3.4. We explicitly want to mention, that the noise adaptation capability of our algorithm described in section 3.1 is absolutely essential. Our experiments have shown that without the adaptive noise threshold almost no events are detected at all. The obtained results are summarized in Table 1. No is the running index of the video, GT is the number of actual freeze frame impairments in the video, TP are true positives, FP denotes false positives, FN are false negative detected freeze frame events and FDR is the false detection rate per frame. Σ denotes the cumulated results over all videos. Note that in all our evaluations (unless stated differently) we use a minimum freeze frame length ($l_{min}$) of 4.

**Table 1:** Basic evaluation performance for all videos. The mean recall ($r_\mu$) and mean precision ($p_\mu$) values are $r_\mu$ = 0.96 and $p_\mu$ = 0.08 respectively.

| No | Duration | GT | TP | FP | FN | FDR |
|---|---|---|---|---|---|---|
| 1 | 0:13:57 | 0 | 0 | 206 | 0 | 0.061 |
| 2 | 0:21:54 | 2 | 0 | 1 | 2 | 0.002 |
| 3 | 0:54:59 | 2 | 2 | 197 | 0 | 0.027 |
| 4 | 0:30:47 | 38 | 38 | 78 | 0 | 0.042 |
| 5 | 0:06:23 | 0 | 0 | 11 | 0 | 0.011 |
| 6 | 0:03:16 | 1 | 1 | 0 | 0 | 0.000 |
| 7 | 0:05:15 | 1 | 1 | 2 | 0 | 0.001 |
| 8 | 0:02:14 | 1 | 1 | 10 | 0 | 0.020 |
| Σ | 2:18:45 | 45 | 43 | 505 | 2 | 0.028 |

As one can see, our algorithm reliably identifies almost all of the annotated freeze frame segments in the videos, but the number of false detections is rather high. A more detailed investigation of the respective events shows, that most of these events are short parts within nearly static or low-motion scenes. This is a motivation for the modeling of human perception as described in section 3.5. It is obvious that if there is no significant motion in the temporal neighborhood humans rather tolerate freeze frame events than in high activity scenes.

### 4.2 Approximation of human perception

The approximation of human perception as described in section 3.5 is a critical point in order to come closer to the defect interpretation similar to the human being. Table 2 shows the influence of this post processing. As one can see, the number of false positives is reduced by a factor of 10, while the number of missed events becomes only slightly higher.

**Table 2:** Results for taking into account the approximation of human perception. The mean recall and mean precision values are $r_\mu$ = 0.89 and $p_\mu$ = 0.45 respectively.

| No | Duration | GT | TP | FP | FN | FDR |
|---|---|---|---|---|---|---|
| 1 | 0:13:57 | 0 | 0 | 3 | 0 | 0.001 |
| 2 | 0:21:54 | 2 | 0 | 0 | 2 | 0.000 |
| 3 | 0:54:59 | 2 | 2 | 36 | 0 | 0.010 |
| 4 | 0:30:47 | 38 | 38 | 3 | 0 | 0.001 |
| 5 | 0:06:23 | 0 | 0 | 1 | 0 | 0.000 |
| 6 | 0:03:16 | 1 | 0 | 0 | 1 | 0.000 |
| 7 | 0:05:15 | 1 | 0 | 2 | 1 | 0.001 |
| 8 | 0:02:14 | 1 | 0 | 3 | 1 | 0.004 |
| Σ | 2:18:45 | 45 | 40 | 48 | 5 | 0.005 |

A detailed investigation of erroneous events shows, that the remaining false positives are mainly due to short static events on the end of a shot (e.g. showing a product image at the end of an advertisement), while missed detections are typically entire freeze frame shots. For reducing these errors human like scene interpretation capability is required to classify such events correctly.

### 4.3 Parameterization and runtime

In order to optimize the parameter settings we extensively evaluated our algorithm using different parameterizations. Due to the lack of space it is not possible to present all the details, but it turns out that the critical parameter is the minimal freeze frame length $l_{min}$. Table 3 shows the dependency of various performance measures on a change of $l_{min}$ for the weather surveillance camera transmissions (video 4, GT denotes the ground truth). While the detection of correct freeze frame impairments (those annotated in the GT) does not suffer, the number of false positives notably decreases for increasing $l_{min}$.

**Table 3:** Dependency of various performance measures on a change of $l_{min}$ for the weather surveillance camera transmissions.

| $l_{min}$ | GT | TP | FP | FN | FDR |
|---|---|---|---|---|---|
| 2 | 38 | 21 | 78 | 17 | 0.005 |
| 3 | 38 | 38 | 6 | 0 | 0.002 |
| 4 | 38 | 38 | 3 | 0 | 0.001 |
| 5 | 38 | 38 | 4 | 0 | 0.002 |
| 6 | 38 | 37 | 3 | 1 | 0.001 |
| 7 | 37 | 37 | 3 | 0 | 0.001 |
| 8 | 37 | 37 | 3 | 0 | 0.001 |
| 9 | 37 | 37 | 3 | 0 | 0.001 |

Figure 1 depicts the dependency of precision and recall rates on a change of $l_{min}$ for the weather surveillance camera transmissions. As most of the false positives originate from short sequences it is comprehensible, that increasing $l_{min}$ leads to better precision and more stable result. So in a final application the user must chose $l_{min}$ according to the individual tradeoff between the ability to detect short freeze frame impairments and the amount of tolerable false alarms.

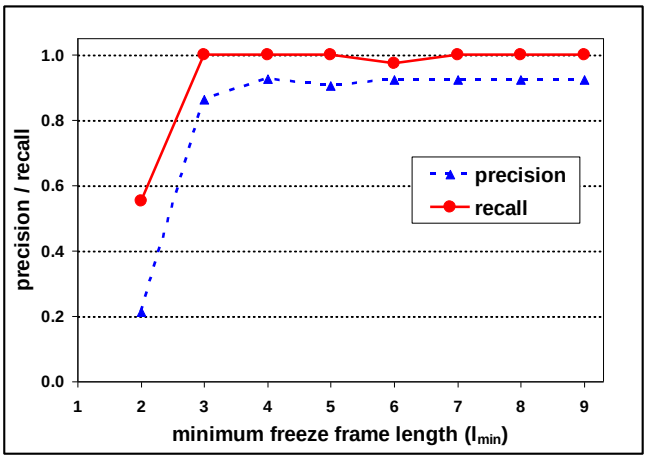


**Figure 1:** Dependency of precision and recall rates on a change of $l_{min}$ for the weather surveillance camera transmissions.

For a fully automatic analysis (high throughput requirement) the estimated runtime of the algorithm is a critical issue. We profiled our code and found most of the time is spent for calculating the activity measures, while the other components of the algorithm (see section 3) are negligible. The overall runtime of the algorithm, without de-coding the individual frames, is between 5ms and 10ms, depending on the number of freeze frames within the content) running on one core of a standard Quad-Core Pentium IV (3GHz, 4GB) for standard definition video (720*576 pixel). Practically it is possible to analyze video significantly faster than real-time.

## 5. APPLICATION

In order to facilitate interoperability and exchange of impairment descriptions between different applications, a standardized way of description is needed. In previous work [1] we have proposed a framework for the description of visual impairments based on MPEG-7 [3]. Based on the existing work in the audio part, we have defined a similar description framework for the visual domain with even more capabilities for describing details of impairments. A list of defects (e.g. freeze frames) and quality measures (e.g. noise/grain level) can be described for each segment, either using a generic descriptor identifying the impairment using a classification scheme or a specific descriptor for a certain impairment type. We have defined a comprehensive impairment classification scheme that provides for hierarchical organization and multilingual description of impairments. The main organization criteria of the classification scheme are the visible and audible effects of impairments. The proposed MPEG-7 extension and classification scheme are available at [4].

In order to enable efficient human interaction with quality analysis results we have proposed in [1] the *Quality Summary* viewer. The temporally condensed overview allows the user to quickly grasp the frequency and strengths of the impairments in the material. The tool supports the user in efficiently navigating the content by providing a timeline representation of a number of views. All views are synchronized with the video player. The temporal resolution can be changed so that the user can freely change the level of detail shown. This tool has been extended in order to allow visualization and direct navigation to freeze frame impairments found within the video, see Figure 2.

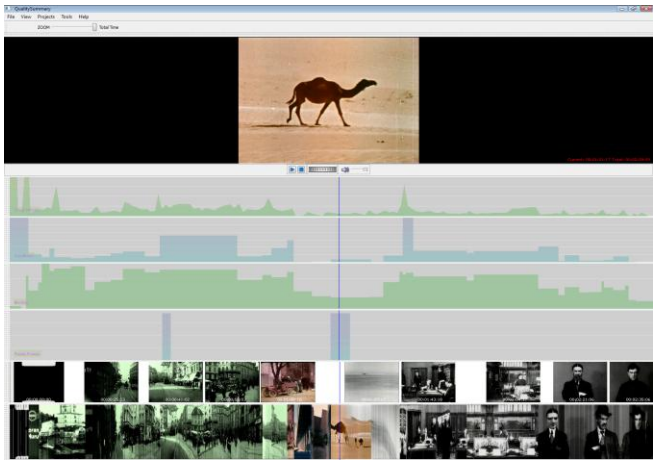

**Figure 2:** The *Quality Summary* viewer.

## 6. SUMMARY AND OUTLOOK

In this paper we have discussed the requirements for automatic quality analysis algorithms in the context of video preservation and presented an algorithm for the detection of freeze frame impairments meeting these requirements. The algorithm evaluation on a set of videos has shown that robustness against other impairments (noise and flickering) is essential. The proposed algorithm allows successful detection of freeze frame impairments with a minimum length of three frames. It has also been shown that the consideration of human perception is important to achieve low false detection rates. The algorithm runs about 5 times faster than real-time for standard definition video on standard PC hardware, assuring high throughput. Future work aims to develop algorithms which interpret the scene on a higher level in order to automatically detect static content like captions or full screen logos.

## 7. ACKNOWLEDGEMENTS

The authors would like to thank Werner Haas, Georg Thallinger, Hermann Fürntratt and Albert Hofmann as well as several other colleagues at JOANNEUM RESEARCH and Martin Altmanninger and Paul Leitner from media services GmbH, who contributed valuable input to the work. This work has been funded partially under the 7th Framework Programme of the European Union within the IST project "PrestoPRIME" (IST FP7 231161) and under the FIT-IT Programme of the Austrian Federal Ministry for Transport, Innovation and Technology within the project “vdQA”.

## 8. REFERENCES

[1] P. Schallauer, W. Bailer, R. Mörzinger, H. Fürntratt and G. Thallinger, “Automatic Quality Analysis for Film and Video Restoration,” *IEEE ICIP*, San Antonio, TX, USA, Oct. 2007.

[2] J. Kim, R. Park, “A Fast Feature-Based Block Matching algorithm using Integral Projections” *IEEE Journal on selected Areas in Communications*, Vol. 10, No. 5, 1992.

[3] ISO/IEC, Multimedia Content Description Interface, ISO/IEC 15938:2001.

[4] *Audiovisual Defect and Quality Description.* URL: http://mpeg7.joanneum.at.