VIDEO SYNCHRONIZATION USING BIT RATE PROFILES

G. Schroth, F. Schweiger*, M. Eichhorn*, E. Steinbach*, M. Fahrmair†, W. Kellerer†*

*Institute for Media Technology, Technische Universität München
†DOCOMO Euro-Labs Munich

ABSTRACT

We present a novel approach for the temporal synchronization of multiple videos which is based on cross-correlating bit rate profiles. The proposed scheme determines the temporal offset without major restrictions on viewing angles, camera properties and camera motion. We propose two extensions of the basic algorithm which reduce the influence of camera motion and distracting background objects. Additionally, we describe how to optimally combine different bit rate components in order to further improve the reliability of the synchronization scheme. The proposed approach, when combined with the three extensions, leads to a reliable, robust, and frame accurate temporal alignment of videos at remarkably low complexity.

***Index Terms*—** Synchronization, video signal processing, video coding

1. INTRODUCTION

Video synchronization is a fundamental requirement for almost all applications involving multiple videos of the same scene. In professional video or TV production, hardware-based synchronization is used to achieve reliable and accurate temporal alignment. Example applications include, *e.g.*, modern sport event TV-transmissions where highlights can be synchronously observed from different viewpoints. The ability to reliably synchronize multiple videos without dedicated hardware would not only imply an enhancement to professional video production but in particular facilitate new community based services. With the increasing availability of camera phones, the probability that a specific event has been recorded by more than one person is very high. Multiple independent recordings allow us not only to change the viewing angle during playback, but also to reconstruct the 3D scene for augmented reality applications, subtract background or foreground objects, or to generate a super resolution video [1]. Hence, a multitude of media enrichments can be provided via collaborative video recording if the individual videos can be reliably synchronized.

1.1. Related Work

Several approaches for software-based video synchronization have been proposed in the last years. A possible classification, published in [2], distinguishes between *feature based* [3, 4, 5], *intensity based* [6, 7], and *camera movement based* algorithms. The latter category comprises very specialized scenarios involving rigidly linked cameras. Feature-based approaches represent the largest family of synchronization methods. Here, it is assumed that image features, *i.e.*, highly discriminative points or edges, can be detected in one video and related to corresponding features in the other sequence. The basic idea is that the motion or occurrence of features, which belong to the same 3D point, are correlated among the different cameras. Tuytelaars and Van Gool [4] have presented a feature-based algorithm that finds the temporal offset by examining the distance of rays of sight. Yan and Pollefeys proposed an approach that bases upon the temporal distribution of space-time interest points [5]. The major disadvantage of this class of algorithms is that reliable detection, matching, and tracking of features through the sequences is required. This non-trivial problem has not been solved satisfactorily yet. Also, camera viewing directions differing by more than 30° are usually not supported.

Intensity-based synchronization algorithms focus on establishing a mapping between the pixels in one video and the pixels in the second one. Caspi and Irani's work on spatio-temporal alignment laid the foundation in this domain [6]. In some approaches not only the temporal offset between two videos is estimated but also the geometric distortion between the two images. Similar to feature-based approaches, intensity-based algorithms are computationally demanding, and also exhibit similar viewpoint constraints.

In [7], general brightness variations in a video are captured by simply summing up temporal intensity derivatives in every frame. This generates a "brightness change profile" over time, which can be compared to that of other cameras observing the same scene. This algorithm requires static cameras and does not tolerate the slightest shaking motion.

For the special case of videos capturing a scene with active sound sources in sufficient quality, audio-based approaches can also be considered for synchronization, as for instance proposed in [8].

To enable the reliable synchronization of videos, we have to overcome three major challenges. First, the computational complexity of feature extraction and of pixel wise processing in intensity-based approaches is prohibitive for large scale applications. Second, the restriction of the viewing angle between cameras is especially unfavorable since many applications actually aim at significant view point changes (*e.g.*, cooperative video in [9]). Finally, a universally applicable synchronization algorithm must cope with camera motion as in many scenarios videos are recorded using hand-held devices.

The remainder of the paper is organized as follows. In Section 2 we introduce our novel, bit rate profile based approach for video synchronization. Three major extensions to the basic approach are presented in Section 3. In Section 4, we evaluate our approach with respect to its invariance to varying view points and camera motion. Section 5 concludes the paper with an outlook to future work.

2. CROSS-CORRELATION OF BIT RATE PROFILES

Unlike previous approaches, which try to imitate the human way of detecting temporal mismatches, we utilize a fundamentally different and high-level fingerprint to align the videos in the temporal domain. The conditional entropy of video frames quantifies the remaining information given the information in previous frames and thus describes the amount of unexpected change in the scene. Under the assumption that the amount of scene variation (caused by

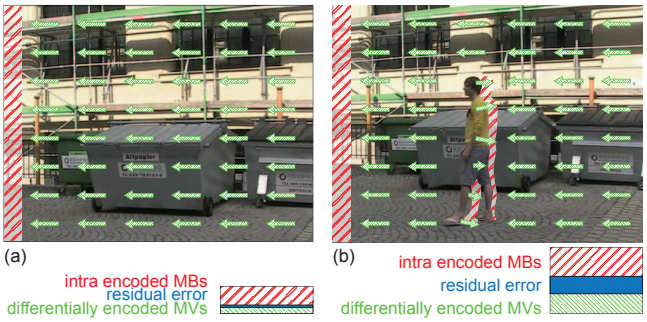

Fig. 1: Simplified qualitative view on bit rate contributions in the cases of (a) sheer camera motion and (b) additional scene changes

dynamic objects) changes over time, the sequence of this measure can serve as a characteristic fingerprint for the video. Further, in realistic scenarios scene alternations can be expected to be visible from almost all viewpoints resulting in correlated conditional entropy profiles. Thus, the temporal offset between two videos can be determined by performing *normalized cross-correlation (NCC)* on their conditional entropy sequences as given in Eq. 1. Here, the $X_i(t)$ are the entropy sequences. $\bar{X}_i$ and σ_{X_i} represent their mean and standard deviation in the overlap region over which the given sum is computed. The normalization makes the correlation invariant to the properties of the cameras, *e.g.*, their resolution. As we do not utilize a geometric measure, no further adaptation is required.

$$C(\Delta t) = \sum_t \frac{[X_1(t+\Delta t) - \bar{X}_1][X_2(t) - \bar{X}_2]}{\sigma_{X_1}\sigma_{X_2}} \qquad (1)$$

Since the conditional entropy of a video sequence is a theoretical measure and depends very much on how the memory of the source is exploited, an approximation has to be used. The bit rate obtained when encoding a video with a standard video codec at fixed quantization is actually indicating the amount of novel information per frame. This approximation further has the advantage that most videos are already encoded and thus hardly any further processing is required to perform the proposed synchronization.

Nevertheless, one might argue that changes in the video and thus the bit rate are not only caused by changes in the scene but also due to camera motion. However, as state-of-the-art codecs like H.264/AVC [10] offer sophisticated motion compensation, the image changes caused by camera motion can be represented with a lower bit rate than complex alternations as they are typical for scene changes. The individual image macro blocks (MB) of the current frame are predicted from preceding frames by determining corresponding macro blocks and their relative translation, represented by motion vectors (MV). A camera pan results in a smooth MV field (see Fig. 1a) which can be efficiently compressed using spatial prediction and differential encoding. Because of the pure displacement of MBs the prediction achieves a small residual error. The area marked in bold red stripes in Figure 1a, which has not been visible in previous frames, adds to the overall data rate in the form of intra encoded MBs.

In contrast, the MVs arising from scene motion carry more information and thus result in a significantly higher bit rate (Fig. 1b). Additionally, as changes in the scene are complex and finely structured, the segmentation during motion estimation typically leads to smaller sub blocks and hence higher rates for block pattern and motion vector signaling. In general, the less effective prediction of the fine and complex image alternations leads to more important residual signals after motion-compensated prediction (MCP) that also have to be encoded and transmitted. Finally, the moving person in Figure 1b, uncovers parts of the background which increases the intra-block contribution as well. Hence, changes in the scene, and thus the entropy or information content of the video, are closely related to the bit rate required to encode it.

With the bit rate as a fingerprint, we can synchronize two videos by cross-correlating their bit rate curves over time, searching for the maximum correlation value. This correlation-based approach is unsusceptible to camera and irrelevant background motion as long as these are uncorrelated between the two videos.

As will be demonstrated in Section 4, this basic algorithm allows for frame-exact synchronization of videos. In the following section, we will introduce three extensions to further increase robustness.

3. EXTENSIONS

With the extensions described in this section we achieve increased robustness against distracting camera and background motion, and additionally exploit the redundancy in the different bit rate components.

Camera motion The major bit rate contribution caused by camera motion is due to newly emerging image content at the frame borders (see Fig. 1). Since in typical scenarios the bit rate in these peripheral areas is almost exclusively induced by camera motion, we exclude their contribution from the considered bit rate profiles to emphasize the actual foreground object motion. To this end, we reduce the effective frame dimensions by ignoring the outer MB fringe. In the following, we will refer to this technique as *border removal*. A similar approach involves macro blocks adjacent to the discussed frame border. This "second ring" also mainly reflects camera motion, but does not so much undergo typical border effects caused by newly emerging areas. Hence, we can estimate the camera motion influence on the entire frame by extrapolating the bit rate from these representative blocks. If a ratio of $b\%$ of the total number of pixels in the frame are covered by the second ring MBs, we take $\frac{100}{b}$ of their bit rate contribution as an estimate for the overall share of camera motion and subtract it from the bit rate profile. In Section 4, we will show how *border removal* and *camera motion extrapolation* can help to improve the synchronization results in the case of non-stationary cameras.

Distracting background While uncorrelated background motion, such as observed by diametrically opposed cameras, naturally does not have significant impact on bit rate correlation, effects which are correlated but temporally unrelated can result in false hypotheses with respect to the temporal offset between the two videos. This can, *e.g.*, occur if a car passes behind the scene and appears some frames earlier in the field of view of one of the cameras due to different viewing directions. If the distracting object passes in front of the scene, the effect is even stronger as the respective bit rate peak outweighs that of the relevant scene objects. To attenuate disproportionately large changes in the bit rate, hence to give equal weight to smaller variances, we propose an additional *companding* of the bit rate profiles prior to correlation. Applying $\arctan(\cdot/c)$ to the median centered and standard deviation normalized signals allows us to shift the sensitivity of the correlation from large peaks only to all significant scene changes. The compression factor c proves to have only little influence and should be set somewhere below 1 bit. Even in the limit $c \to 0$, where the compression degenerates to a mere binary decision, acceptable synchronization remains possible.

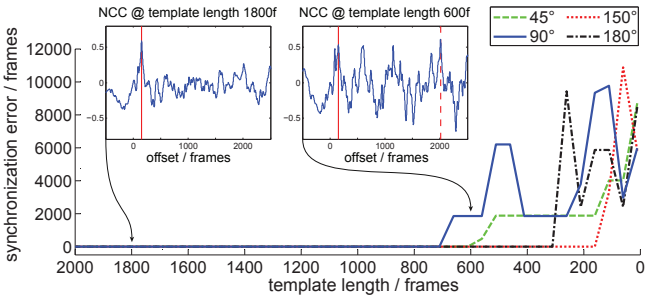

Fig. 2: Synchronization offset error as a function of the length of the template taken from the respective second sequence. Four sequences were compared to the same reference differing from it in viewing direction (from 45° to 180°). The two embedded graphs show the normalized cross correlation functions for the view at 90° at exemplary template lengths. The solid red line indicates the ground truth offset, the dashed one an erroneous synchronization result.

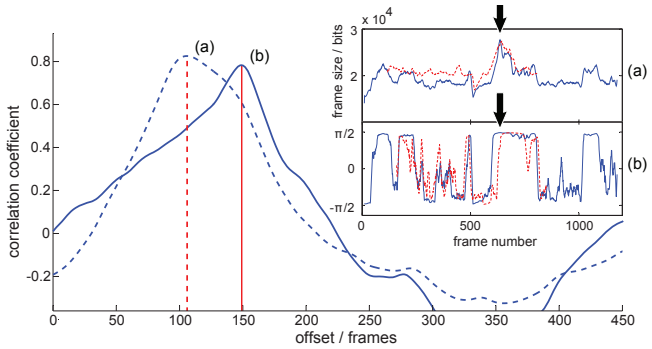

Fig. 3: Correlation result without (a) and with companding (b). The solid red vertical line in the main plot identifies the ground truth offset between the two videos. (a) Without companding, the NCC snaps in at the misleading bit rate peak marked with an arrow in the upper embedded figure. (b) If the bit rate sequences undergo compression with $c = 0.1$ bit, this peak no longer dominates and is overruled by the majority of smaller yet consistent bit rate changes.

Information fusion As described in Section 2, the bit rate contains contributions from the MVs and "texture data" (comprising the residual error after MCP and independently encoded MBs). However, due to the significantly higher bit rate required to encode the latter, the information contained in the MVs hardly influences the cross correlation result. Further, it is important to note that for coarse quantization, the temporal fingerprint in the MV data is less pronounced, whereas for fine quantization the variation of the texture bit rate is very limited. Thus, both MV and texture bit rate have to be correlated separately, and the results C_{MV} and C_{TEX} combined in an optimal manner to exploit all given information and to be substantially independent in the choice of the quantization parameter. When the difference of the perfectly aligned bit rate profiles is modeled as additive Gaussian white noise, according to [11], the maximum likelihood (ML) combination of C_{MV} and C_{TEX} is given by Eq. (2). Obviously, this calculation rule cannot innately handle antipodal input signals. However, since both correlation functions C_{MV} and C_{TEX} should take positive values at the true temporal offset, negating C_{ML} whenever any of the two is negative resolves the ambiguity in practice.

$$C^2_{\text{ML}}(\Delta t) = 1 - \sqrt{[1 - C^2_{\text{MV}}(\Delta t)]\,[1 - C^2_{\text{TEX}}(\Delta t)]} \qquad (2)$$

4. EXPERIMENTAL RESULTS

Throughout this paper, we use the H.264 codec to generate the bit rate traces for our experiments[1]. In particular, we modified the code of FFmpeg and the x264 library such as to dump information on macro block sizes and their types into a file. For all experiments described here, we disabled bidirectional prediction and enforced a GOP composed of 500 frames (IP...P) . The bit-rate values for the I-frames were removed from the bit rate sequence by simply replacing their value with the one of a neighboring frame. The only parameter that we found to be crucial is the quantizer level q used by H.264. For values ranging from $q = 20$ to $q = 40$, we obtained the best results. Despite this coarse quantization, the original input GOP structure (IPP in our case) was still visible after re-encoding. Smoothing the bit rate curves with a moving average filter of width 10 had a positive influence in alleviating this artificial periodicity.

Required template size A general problem when evaluating the cross-correlation of two sequences is the range of actually examinable lags. The length difference of the two sequences determines the number of shifts for which the shorter sequence fully overlaps with the longer one. Whenever the sequences interleave only partially, the computed correlation value might be flawed if the overlap is much smaller than the sequence lengths. We rather conservatively decided to consider only those lags as valid which permit complete overlap. To still be able to examine a reasonable range of potential offsets, it is hence necessary to use only an excerpt from the shorter sequence. Consequently, a trade-off has to be made between correlation reliability (using a long excerpt) and the possible range of examinable offsets (given a short excerpt).

In a first experiment, we applied the basic version of our algorithm to videos of a person in steady motion recorded by 5 tripod mounted *Canon FS100* camcorders located on a semi-circle of radius 4 m around the scene. Each video is about 8 minutes long, interlaced, and in PAL resolution (720×576) at 25 fps. We provided the subject with a foot-bag and instructed him to play without leaving the cameras' field of view. The respective ground truth synchronization offsets were determined by visual inspection. Fig. 2 shows the dependency of successful synchronization on the length of the used excerpt or template. The main plot shows 4 graphs corresponding to camera pairs with the given viewpoint differences. For template lengths above 700 frames, the synchronization is frame-accurate in all cases. This value is of course scene specific; we found that for more distinct motion patterns, excerpts as short as 100 frames can be sufficient. This experiment also shows that our algorithm is applicable irrespective of the camera pair's viewing directions. If at all, it is susceptible to cameras observing the scene from perpendicular directions. Accordingly, the 90° sequence requires the longest minimum template length to be successfully synchronized with the 0° reference.

Companding The next experiment illustrates how compressing the bit rate signals prior to correlation can alleviate the effects of spurious distracting background motion. To this end, we captured the actions of a person using two tripod mounted camcorders.

[1]The videos used in our experiments can be downloaded from `http://www.lmt.ei.tum.de/florian/sync/`

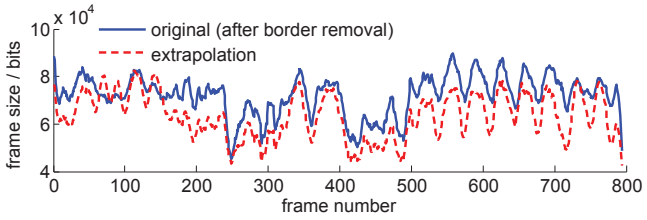

Fig. 4: In case of pure camera motion, the bit rate extrapolation closely follows the original bit rate.

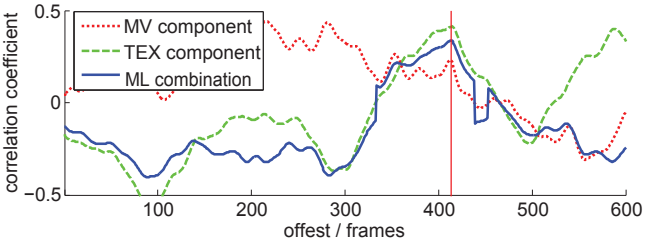

Fig. 5: NCC of the individual components fails to recover the true offset (indicated by the vertical line). The ML combination of the two curves yields a conclusive peak at the correct position.

All parameters were chosen identical to those in the template size experiment, the cameras' viewing directions formed an angle of roughly 70°. We instructed the subject to walk into the scene, hide and reappear before leaving the scene entirely. The resulting bit rate sequences exhibit a distinct motion pattern that, by itself, would be easily matchable. Around frame number 600 (*cf.* Fig. 3a), however, a group of passers-by occlude the happening causing a short-time bit rate increase. Walking in from the right, their impact occurs first in the right view, and is not visible by the left camera but 43 frames later. Since the interferer bit rate peak is stronger than the contribution of the performer in focus, NCC tends to misalign the sequences with an error of exactly 43 frames. Figure 3 illustrates how companding is beneficial in this case, aiding the NCC to find the correct offset despite the disturbance.

Border removal and camera motion extrapolation To demonstrate the effectiveness of the proposed camera motion compensation scheme we apply it to a video recorded with a handheld camera in front of a static background. In addition to the inevitable freehand shaking, we deliberately performed slight panning motions. Since there are no dynamic scene parts, the resulting bit rates are fully determined by the camera motion. Hence, a combination of border removal and camera motion extrapolation should ideally be able to reproduce the observed bit rate variations. In Figure 4, the original bit rate sequence after border removal, *i.e.*, with the rate of the outermost MBs deducted, is compared to the extrapolation based on the second outermost MBs. Apparently, the involved effects can be well emulated with data from the peripheral areas only. Camera motion can thus be substantially removed from a given video.

ML combination of bit rate components In this last experiment, we consider a combination of all the effects discussed so far. Our scene is composed of a person frisking and waving at the cameras, which are some 50° apart this time. The background is dynamic, again there is a second person passing by the cameras, and one of the cameras is handheld introducing significant shaking and panning motion. We apply all the proposed extensions at a time in order to achieve frame-exact synchronization despite the unfavorable conditions. Figure 5 exemplarily shows the correlation outcome for a *non-ideal* choice of the template excerpt. Here, the correlation of the individual parts of the bit rate, namely MV and texture bits (both after border removal/extrapolation and companding), are plotted separately. Obviously, both curves show a peak at the desired offset, but there are other lags with higher correlation values. The joint NCC curve, however, obtained following Eq. (2), exhibits a global maximum at the correct position. Offsets falsely favored by one of the components are rejected by the other and vice versa.

5. CONCLUSION

In this paper, we have presented a novel approach to video synchronization. To the best of our knowledge it is the first algorithm capable of synchronizing videos recorded with unconstrained viewing directions, unknown camera properties, and in the presence of significant camera motion. In conjunction with its low computational complexity it allows to provide autonomous video synchronization to mass market applications. The maximum likelihood combination of the information given in the texture and MV bit rate sequences facilitates a frame accurate synchronization even for very challenging scenarios. Further, the proposed extensions provide additional robustness against camera motion and misleading temporal offset hypotheses. Upcoming enhancements will include the ability to cope with varying frame rates and robustness against misleading scene alternations. Also, we will investigate the synchronization of videos that have been encoded using different video codec standards.

6. REFERENCES

[1] E. Shechtman, Y. Caspi, and M. Irani, "Increasing space-time resolution in video," *Lecture Notes in Computer Science*, vol. 2350, pp. 753–768, 2002.

[2] C. Lei and Y. H. Yang, "Tri-focal tensor-based multiple video synchronization with subframe optimization," *IEEE Trans. on Image Processing*, vol. 15, no. 9, pp. 2473–2480, Sept. 2006.

[3] G.P. Stein, "Tracking from multiple view points: Self-calibration of space and time," in *CVPR*, Ft. Collins, CO, USA, June 1999.

[4] T. Tuytelaars and L. Van Gool, "Synchronizing video sequences," in *CVPR*, Washington, DC, USA, June 2004.

[5] J. Yan and M. Pollefeys, "Video synchronization via space-time interest point distribution," in *ACIVS*, Brussels, Belgium, Sept. 2004.

[6] Y. Caspi and M. Irani, "A step towards sequence-to-sequence alignment," in *CVPR*, Hilton Head, SC, USA, June 2000.

[7] M. Ushizaki, T. Okatani, and K. Deguchi, "Video synchronization based on co-occurrence of appearance changes in video sequences," in *ICPR*, Hong Kong, Aug. 2006.

[8] P. Shrestha, M. Barbieri, H. Weda, and D. Sekulovski, "Synchronization of multiple camera videos using audio-visual features," *IEEE Trans. on Multimedia*, vol. 12, no. 1, pp. 79–92, Jan. 2010.

[9] F. Schweiger, E. Steinbach, M. Fahrmair, and W. Kellerer, "CAMP: A framework for cooperation among mobile prosumers," in *ICME*, New York, NY, USA, June 2009.

[10] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H. 264/AVC video coding standard," *IEEE Trans. on circuits and systems for video technology*, vol. 13, no. 7, pp. 560–576, 2003.

[11] S. Zucker, "Cross-correlation and maximum likelihood analysis: a new approach to combine cross-correlation functions," *Monthly Notices of the Royal Astronomical Society*, vol. 342, no. 4, pp. 1291–1298, July 2003.