Cross-Modal Analysis of Audio-Visual Film Montage

Matthias Zeppelzauer, Dalibor Mitrovic, Christian Breiteneder
Vienna University of Technology
Interactive Media Systems Group
Favoritenstrasse 9-11, 1040 Vienna, Austria
Email: lastname@ims.tuwien.ac.at

***Abstract*—A stylistic device frequently employed by filmmakers is the *synchronous montage* (composition) of audio and visual elements. Synchronous montage helps to increase tension and tempo in a scene and highlights important events in the story. Sequences with synchronous montage usually contain rich semantics which is relevant for understanding a movie. This property is currently not exploited in automated indexing, annotation, and summarization of movies. We propose a cross-modal approach that extracts sequences from a movie with synchronous audio-visual montage. Experiments confirm that the extracted sequences have high semantic relevance. Consequently, they represent a useful basis for different high-level movie abstraction tasks such as automated movie annotation and movie summarization.**

I. Introduction

Film montage (also known as film editing) addresses among others the composition of audio and visual elements for the purpose of story telling. A well-established technique for audio-visual composition is the *synchronous montage*. Synchronous montage relates to the synchronicity between events in the soundtrack (e.g. a sudden noise) and the cutting of the movie (e.g. a shot cut). Note that this synchronicity is at a different (higher) structural level than e.g. lip synchronicity (which is not addressed in this work). Synchronous audio-visual montage enables the director to accentuate important events and actions and to increase tension and tempo (e.g. in action scenes and dialogue sequences) [1]. Such sequences contain rich semantic context which is important for understanding a movie.

A famous example for the synchronous montage technique in film history stems from the film “Enthusiasm” by Dziga Vertov from 1931. “Enthusiasm” is a Soviet propaganda film about the first Soviet five-year plan. A central sequence in the film shows several consecutive static shots of different religious and monarchal symbols (e.g. a tsarist monogram, statues of Christ, crucifixes). At each shot boundary between two different symbols the director positioned the powerful sound of a church bell in the soundtrack. The synchronous church bells increase the perceptual salience of the sequence and create a threatening and warning atmosphere. According to the film literature, this is a key scene in the film that expresses the rejection of religion and the tsarist regime by the communist regime [2]. An excerpt of the sequence together with the waveform of its soundtrack is shown in Figure 1.

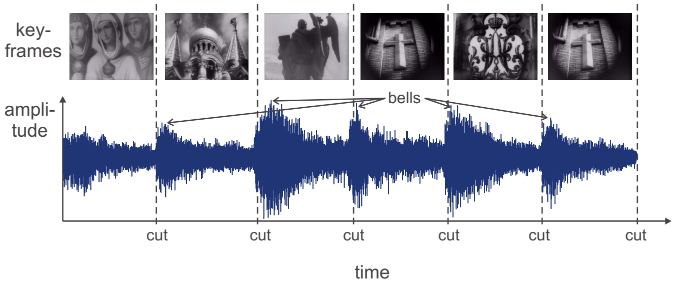


Fig. 1. A synchronous montage sequence. The keyframes of each shot show different religious symbols. The peaks in the waveform’s amplitude at the shot cuts correspond to the church bells.

Synchronous montage is still a popular technique in contemporary movies to emphasize important events (a detailed discussion is provided in [1]). For example, in the movie “The Hunt for Red October” from 1990 the director exploits synchronous montage in dialogue sequences to emphasize the speakers and the speech. Another example mentioned in [1] is the end scene (the showdown) of “The Last of the Mohicans” where the cutting of the movie is coordinated with the musical rhythm.

Due to their rich semantics, synchronous montage sequences are important for (automated) movie annotation, interpretation, and summarization. For example, synchronous montage sequences are likely to contain key scenes which should be part of an automated generated movie summary or trailer. Additionally, film professionals are interested in the extraction of such sequences for film and montage analysis.

In this paper, we investigate the automated extraction of sequences with synchronous audio-visual montage. For this purpose we develop a cross-modal approach that extracts such sequences by detecting temporally correlating audio and visual events. Unfortunately, the temporal correlation[1] of audio and video on the signal-level differs significantly from the correlation on the perceptual level. Consequently, established methods for the estimation of temporal correlations do not work properly. We propose an approach for the extraction of temporal correlations that are more meaningful and intuitive for the human observer. First, we extract salient audio and

[1] Note that “correlation” in this work is not meant in a strict statistical sense.

visual events by the detection of *onsets*. In general, onsets represent abrupt changes in the underlying signals. Visual onsets refer to abrupt shot boundaries (cuts). In the audio domain, onsets are for example musical beats, sudden sound effects, and points in time when an actor starts to speak after a pause. Next, we detect temporally correlated audio and visual onsets by analyzing their coincidences and their temporal neighborhoods. Finally, we extract entire sequences that contain several subsequent correlated audio and visual onsets (synchronous montage sequences). Experiments with different films show that the approach is able to retrieve relevant montage sequences. Additionally, the results include key scenes with rich semantics.

The paper is organized as follows. In Section II we discuss related work. Section III introduces the proposed approach. The experimental setup and results are presented in Sections IV and V. We conclude the paper in Section VI.

II. Related Work

Audio-visual synchronicity (correlation) has been studied by researchers in different domains such as sound source localization [3], [4], talking face detection [5], [6], speech recognition [7], and surveillance [8]. The computation of temporal audio-visual correlation is performed at different *levels*. Most approaches compute correlation directly between audio and visual features (*feature-level*). Frequently employed correlation measures are Pearson correlation [7] and mutual information [5]. Some methods first reduce dimensionality (e.g. by Canonical Correlation Analysis) and then perform correlation computation in a lower-dimensional space [6].

On the feature-level we are able to capture the *natural* correlation that exists between audio and visual signals (from the same source), e.g. speech and lip movements. Consequently, it is well-suited for talking face detection and person identification. However, at this rather low level it is difficult to integrate delays, tolerances, and irregularities into the correlation computation. This limits the applicability of such methods for the analysis of film montage since delays and irregularities are sometimes introduced by the filmmaker for stylistic reasons.

Other methods (especially from the surveillance domain) compute temporal correlations on the basis of classified high-level decisions (*decision-level*) [9]. Methods at this level learn frequent audio and visual events (atomic events) autonomously and recognize higher-level events (e.g. running, opening a door) by merging co-occurring atomic event classifications [8]. Such methods are usually designed to operate in controlled environments (e.g. a corridor in a building) and require recurring events for learning. Both, recurring events and controlled environments are not available in feature films.

For the analysis of audio-visual montage, a method is required that (i) enables flexible temporal correlation assessments and (ii) operates on an uncontrolled (general purpose) set of events. For that reason, we perform the temporal correlation analysis on an intermediate level: the *landmark-level*.

On the landmark-level we operate on salient points (automatically detected peaks and onsets) in the audio and visual feature vectors [10]. This level facilitates the representation of general purpose events and a flexible temporal correlation estimation. Additionally, psychophysical research points out that the human synchrony perception relies on the matching of salient features (peaks and troughs) in the audio and visual modalities [11].

Only little work on audio-visual correlation estimation on the landmark-level exists. Barzelay and Schechner perform sound localization by correlating audio and visual onsets [3]. The audio onsets are derived from a spectrogram and the visual onsets are extracted from motion trajectories. Temporal coincidences of onsets are detected by a likelihood function that yields high values where audio and visual onsets temporally coincide. Similarly, Monaci and Vandergheynst perform sound localization by correlating onsets in audio and visual feature vectors [4]. From the audio and visual onsets the authors first compute two binary vectors where spikes indicate onset positions. Next, they broaden the spikes with a rectangle function to increase the temporal tolerance. Finally, they combine both vectors by a logical AND to obtain temporally correlated audio-visual onsets. The method is for example applied to talking face detection.

The approaches in [3] and [4] are not applicable to the analysis of audio-visual film montage. First, the approaches are designed for correlating sound with *motion*. For the analysis of synchronous montage visual onsets are shot boundaries and not motion. Second, the approaches above consider consecutive onsets as independent from each other and neglect their neighborhood relationships. Thereby, information on the temporal distribution of the onsets is lost which is important to evaluate the salience of an onset. Third, both approaches neglect the actual strengths of the onsets (their degree of abruptness). In fact, the strength is a further indicator for the salience and is important to obtain estimates in accordance with human assessment.

III. Analysis of Synchronous Montage

An overview of the entire approach is depicted in Figure 2. We first perform onset detection in the visual and audio domains separately. Onsets in the visual signal represent shot boundaries. Audio onsets correspond to musical beats, sound effects (e.g. explosions, cries, sirens), and speech onsets. Next, we detect temporally correlating (synchronous) audio and visual onsets (for instance a sudden cry that occurs simultaneously with a shot boundary). Finally, we extract entire sequences that contain several consecutive shot boundaries with correlated audio onsets.

A. Visual Onset Detection

Shots are the most important building blocks of visual film montage. We detect shot boundaries (visual onsets) as described in [12]. First, we extract features for each frame (Edge Histogram, DCT Coefficients). Next, we perform a temporal self-similarity analysis (similarly to [13]) for both

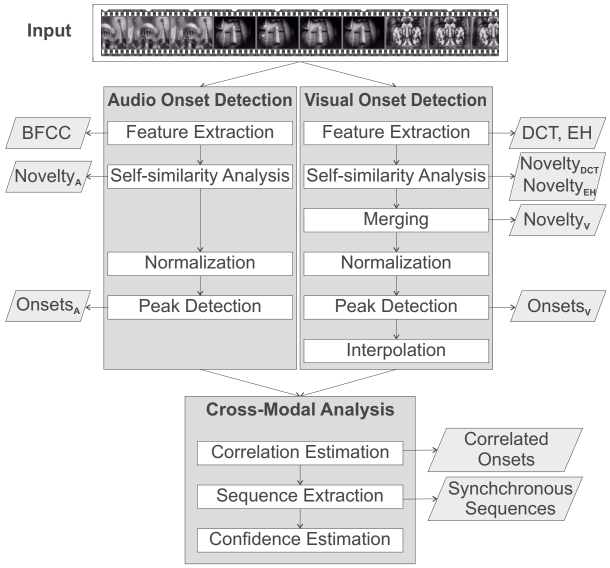


Fig. 2. Overview of the approach.

features and merge the outcomes. The result of self-similarity analysis is a so called *novelty curve* which has peaks at positions where the underlying signal changes abruptly. In the case of the visual signal, peaks in the novelty curve indicate shot boundaries. The shot boundary positions are extracted by a peak detector (after normalization of the novelty curve) and form the set of visual onsets for subsequent processing. Finally, we interpolate the visual onset positions to make them comparable with the audio onsets extracted in the next section. We neglect gradual transitions (e.g. dissolves and fades) since they do not represent distinct events in time that audio onsets can be correlated with.

B. Audio Onset Detection

For audio analysis, we extract 24 Bark-frequency cepstral coefficients (BFCCs) from audio frames of 30ms (20ms overlap). BFCCs employ a psychoacoustically scaled filter bank and compactly represent the coarse spectral frequency distribution in an audio frame. The BFCCs are input to a self-similarity analysis as well (like the visual features in Section III-A). The result is again a novelty curve. In the case of audio, peaks indicate abrupt changes (discontinuities) in the audio stream. Such discontinuities occur for example at the beginning of musical beats, speech, and special effects. The stronger a discontinuity the higher is the amplitude of the peaks. We normalize the novelty curve and extract salient peaks with an adaptive peak detector. The positions and heights of the extracted peaks form the set of audio onsets for subsequent processing.

C. Temporal Audio-Visual Correlation Estimation

The goal of the next step is to find temporally correlated (synchronous) audio and visual onsets which would also be perceived synchronous by a human observer. In general, onsets are perceived correlated if they are temporally near to each other. This conforms with the assumptions made in [3], [4], and [11]. In our case, the correlation of onsets means that an audio onset occurs simultaneously with a shot boundary. However, we observe that this assumption is not sufficient for the detection of synchronous montage in feature films for two reasons. First, stronger (more salient) onsets catch the viewers attention more than weak onsets. Consequently, we integrate a *salience condition* into the correlation computation that favors stronger onsets (originating from higher peaks).

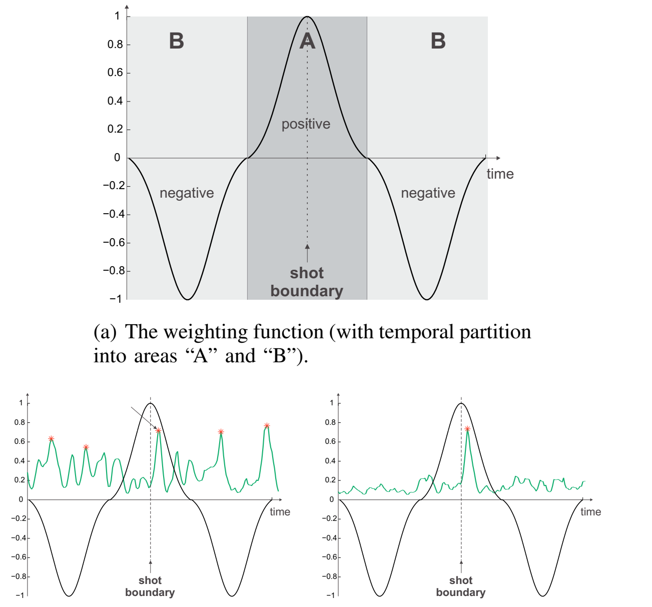
Second, the temporal distribution and the number of audio onsets around a shot boundary influence synchrony perception: if many onsets surround a shot boundary, they distract the attention of the observer from detecting synchronicity. Consequently, a single isolated audio onset at a shot boundary leads to a stronger synchronicity than several audio onsets surrounding a shot boundary (e.g. several overlaid musical beats and background noise). To take this effect into account, we integrate an *isolation condition* into the correlation computation that favors isolated audio onsets.

For temporal correlation estimation we design a weighting function that takes the salience and the isolation condition into account. The weighting function (see Figure 3(a)) is centered around a shot boundary. The amplitude represents the time-dependent influence of an audio onset for temporal correlation estimation. The function can be partitioned into two different areas.

In area “A” centered around the shot boundary the function is positive. Audio onsets that fall within this area influence correlation positively (the nearer the audio onset is to the shot boundary the higher is its influence). The weighting function in area “A” models a simple principle of human synchrony perception: Events that are temporally near to each other are perceived as correlated. With increasing distance the perceived correlation decreases.

In area “B” the function is negative. Onsets that fall into this area get negative weights. If numerous audio onsets (e.g. originating from different background sounds) surround a shot boundary, they contribute negatively to the correlation estimate. An example is shown in Figure 3(b). The peaks in the audio novelty curve marked with an asterisk correspond to detected onsets. Even though the central onset (marked with an arrow) is close to the shot boundary, the overall correlation at this shot boundary is low because the four surrounding onsets have negative weights. This behavior models the isolation condition: the surrounding onsets distract the observer from the central onset which reduces the degree of perceived synchronicity. Figures 3(b) and 3(c) illustrate the effect of the isolation constraint. The shot boundary with the isolated onset in Figure 3(c) yields a higher correlation c_j than the boundary with the surrounded onsets.

The correlation computation is performed as follows. Given a set of audio onsets with positions p_i and heights h_i, $i = 1, ..., M$ and a set of visual onset positions (shot boundaries) b_j, $j = 1, ..., B$, we center the weighting function w around a shot boundary b_j. Note that the weighting function is zero outside of the negative area “B”. The correlation c_j at shot



(a) The weighting function (with temporal partition into areas “A” and “B”).

(b) surrounded onsets: $c_j = -0.1$

(c) isolated onset: $c_j = 0.6$

Fig. 3. The weighting function and examples of positive and negative correlation: the isolated onset yields a higher correlation c_j than a series of onsets that surrounds a shot boundary.

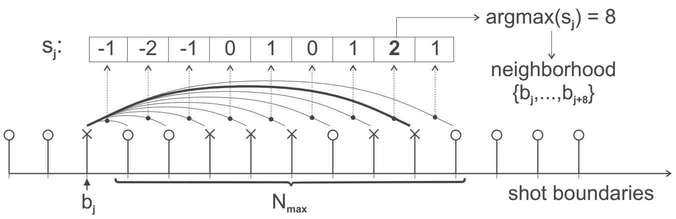


Fig. 4. The schema for the extraction of a neighborhood region at a shot boundary b_j. The maximum sum s_j is obtained for a neighborhood of 8 shots.

boundary b_j is the sum of the products of the weighting function w (at position p_i) with the corresponding heights h_i of the audio onsets: $c_j = \sum_{i=1}^{M} w(p_i) * h_i$.

By taking the actual onset heights h_i into account we are able to model the salience condition. Higher onsets are more distinctive and influence correlation more than lower onsets.

The result is a correlation estimate for each shot boundary. In the following, we consider shot boundaries with correlation $c_j > 0$ as *synchronously edited boundaries* and shot boundaries with $c_j \leq 0$ as *asynchronously edited*.

D. Extraction of Synchronous Montage Sequences

In the synchronous montage technique the director makes *repeated* use of synchronously edited shot boundaries to attract the attention of the viewer over larger time spans. In practice however, such a sequence might contain also some shot boundaries that are purposely *not* synchronized with the audio track by the filmmaker (e.g. for stylistic reasons). For automated sequence extraction on a technical level, we therefore have to search for *possibly interrupted temporal groupings* of synchronously edited shot boundaries.

For this purpose, we propose a tolerant segmentation scheme consisting of two stages. In the first stage, we search for *neighborhood regions* at each synchronously edited shot boundary. The size of the neighborhood regions is maximized on the condition that the number of *irregularities* in the neighborhood (asynchronously edited shot boundaries) is minimized. In the second stage, we merge the (overlapping) neighborhoods to obtain the final montage sequences.

The first stage is illustrated in Figure 4. The extraction of neighborhood regions takes place at synchronously edited shot boundaries (marked with “x” in Figure 4) only. Asynchronously edited boundaries marked with “o” can be skipped. At a given shot boundary b_j we position a support window of size n, where n defines the number of neighboring shot boundaries that are taken into account. Next, we count the number of positively correlated shot boundaries in the support window and subtract the number of negatively correlated boundaries. This results in a sum $s_{j,n}$ for the support window of size n at boundary b_j.

We perform the computation of $s_{j,n}$ for different support windows sizes $n = 1, ..., N_{max}$ which results in a series of sums $s_j = s_{j,1}, ..., s_{j,N_{max}}$ for the shot boundary under consideration (see s_j in Figure 4 for the sums of the example sequence). Next, we estimate for which window size n the maximum sum is obtained: $n_{max} = \underset{n}{\operatorname{argmax}}(s_j)$.

In the example in Figure 4 the maximum sum is obtained for $n = 8$ (sum is 2). Finally, the region from shot boundary b_j to $b_{j+n_{max}}$ is stored as a new neighborhood region. If the maximum sum $s_{j,n_{max}}$ is smaller than 2 no neighborhood region is generated. The process described above is repeated for all synchronously edited shot boundaries. The result is a set of (possibly overlapping) neighborhood regions.

In the second stage, we compute the union of all neighborhood regions in order to obtain the final montage sequences. For each extracted sequence we compute a measure that reflects the confidence in the decision that an extracted sequence actually is a synchronous montage sequence. A straightforward measure is the number of synchronously edited shot boundaries in an extracted sequence. The higher this number the likelier it is that the extracted sequence is a synchronous montage sequence.

IV. Experimental Setup

We evaluate the proposed method with contemporary movies as well as historic film material from the early years of sound film. Especially, the historic material is well-suited for the evaluation of the proposed method because (i) it has low sound and image quality (noise, distortions) and thus allows for the evaluation of the robustness of the method and (ii) the filmmakers of the early sound films intensively experimented with the usage of sound in film montage and as a result the films frequently contain montage sequences with strong audio-visual correlations.

A. Data

The historic material includes the film “Enthusiasm” by Dziga Vertov from 1931 and “October” by Sergei Eisenstein from 1927. Both films stem from Soviet filmmakers who are known for their innovative and experimental montage style. Both films represent different types of soviet propaganda films.

In “Enthusiasm” Vertov deliberately coupled “visible and audible moments” to create a strong tension between sound and visuals. This resulted in a revolutionary style of audio-visual montage [2]. The film “October” from Eisenstein is an (originally silent) fictional film in celebration of the 10th anniversary of the October Revolution. In 1966 a soundtrack containing sound effects and music was added. “October” contains highly formalistic visual montage which partly correlates with the later added soundtrack [14].

The contemporary feature films include “The Hunt for Red October” directed by John McTiernan and “Fight Club” by David Fincher. “The Hunt for Red October” was selected because it is a good example of synchronous montage according to [1]. For “Fight Club” no prior information on the montage style was available. The movie was selected to broaden the test set and to reduce the bias introduced by the other selected movies.

B. Ground Truth

There is no ground truth available for the performed investigation. The consequences for our evaluation are twofold. First, in absence of ground truth we cannot compute recall and precision. Nevertheless, we are able to evaluate the retrieved sequences manually and compute the precision for result sets of different sizes (e.g. for the 3, 5, and 10 sequences with the highest confidence).

Second, we attempt to generate a ground truth for selected material to enable a more comprehensive evaluation. We select “Enthusiasm” which makes the most intensive use of synchronous audio-visual montage. Together with domain experts we annotate synchronously edited shot boundaries and the synchronous montage sequences.

C. Parameters

The correlation computation requires the specification of two parameters: the width of the weighting function w (see Section III-C) and maximum support window size N_{max} (in unit shot boundaries, see Section III-D). We experiment with widths of w of 1 to 1.8 seconds. The wider the function the more temporal tolerance is allowed in the correlation computation. Typical values for N_{max} are between 5 and 11. The larger the values of N_{max} the more irregularities are tolerated during segmentation.

V. Results

We first evaluate the retrieval performance of the proposed method with the generated ground truth. For comparison, we integrate the correlation computation by Monaci et al. [4] into our method. We define two different system configurations: the proposed method with the weighting function as correlation estimator from Section III-C, short: “System P” and as alternative system the proposed method with the correlation estimation of [4], short: “System A”. Both systems operate on the same audio and visual onsets.

TABLE I
Performance of the two compared system configurations evaluated against the ground truth.

	System P			System A		
Task	Rec.	Prec.	F1	Rec.	Prec.	F1
#1 Boundaries	0.67	0.64	0.65	0.88	0.48	0.62
#2 Sequences	0.85	0.72	0.78	0.97	0.41	0.58

Table I presents the results of both systems for the film “Enthusiasm”. We compute recall and precision for two different tasks: first, the detection of synchronously edited shot boundaries (task #1) and second, the extraction of synchronous montage sequences (task #2) which is based on the first task.

The probability for “Enthusiasm” that a shot boundary is synchronously edited is 0.36. The probability of occurrence of a synchronous montage sequence is 0.35. This means that for both tasks random guessing would result in a recall of approximately 0.5 and a precision of approximately 0.36 and 0.35, respectively.

From Table I we observe that System A yields a relatively high recall but a precision which is near random. The reason is that nearly all shot boundaries are classified as “synchronously edited” and during sequence extraction large sequences are extracted that cover nearly the entire film. This is best illustrated in Figure 5 (lower part) which shows the strong under-segmentation produced by the alternative system.

A finer segmentation requires a better balancing between recall and precision (especially a higher precision). System P yields a higher precision and overall F1 measure. This significantly improves the accuracy of the sequence extraction. Again this is best observed from Figure 5 (upper part) where the extracted sequences much better match with the ground truth. Most of the ground truth sequences are partly or entirely retrieved. There are only a few short false positive sequences. The increase of precision is due to the consideration of the isolation and salience condition in the weighting function.

From Table I we further observe that the proposed method (System P) yields higher recall and precision for sequence extraction (task #2) than for single shot boundaries (task #1), although both tasks build upon each other. The reason lies in the tolerance of the segmentation scheme which is able to compensate for falsely classified shot boundaries.

Table II presents the retrieval performance in terms of precision for differently sized result sets (short “P@N” for a result set of size N) for the films from Section IV-A. We obtain the different result sets by retrieving only the N sequences with the highest confidence. Among the first ten retrieved sequences in average 72% are relevant synchronous audio-visual montage sequences. Furthermore, in the film “Fight Club” where no prior information about the montage style was available we discovered several synchronous montage

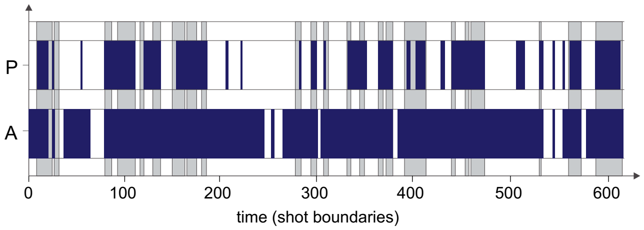


Fig. 5. Sequence extraction results over time (x axis). The regions in the background (gray) represent the sequences in the ground truth. The regions in the foreground (blue) represent the sequences extracted by the proposed system P and alternative system A. While System A generates a strong under-segmentation, System P achieves a finer and more accurate segmentation.

TABLE II
PRECISIONS OF THE PROPOSED METHOD FOR DIFFERENT RESULT SET SIZES (1, 3, 5, AND 10) AND FEATURE FILMS.

Feature Film	P@1	P@3	P@5	P@10
Enthusiasm	1.00	1.00	1.00	0.90
October	1.00	0.67	0.80	0.50
Hunt for Red October	1.00	0.67	0.60	0.70
Fight Club	1.00	0.67	0.80	0.80

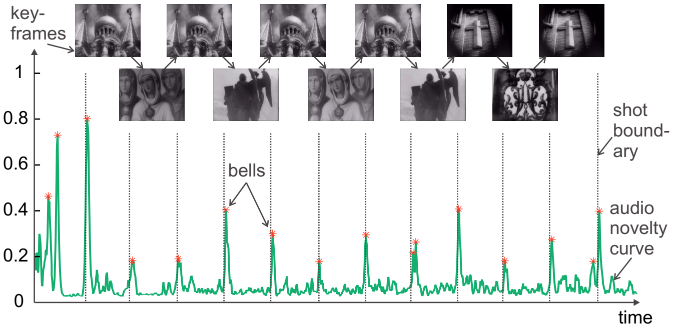


Fig. 6. A sequence showing different religious symbols with synchronously edited bell sounds at each shot cut.

sequences. False positives are returned mostly in situations where a lot of background noise is present in the soundtrack.

From Table II we observe that the performance for historic material is similar to that of the contemporary material. This is remarkable since the historic material contains numerous artifacts in the visual signal (e.g. flicker, shaking, low contrast) as well as in the audio track (e.g. broad-band noise, distortions). There are two reasons for the robustness of the approach. First, we rely on visual and audio onsets which correspond to peaks that are robust to noise to a high degree. Second, even in case of falsely detected onsets, the tolerant segmentation scheme compensates for most of these errors.

The retrieved results include sequences of high semantic interest. For example the top ranked sequence in "Enthusiasm" is the already mentioned sequence of religious symbols from Section I. An illustration of the sequence together with the audio novelty curve is shown in Figure 6. The peaks clearly correspond to the church bells at the shot boundaries.

An interesting observation concerning the sequence in "Enthusiasm" is made from the results for the film "October". One retrieved sequence from "October" shows a similar sequence of religious symbols which are emphasized by bell sounds at each shot boundary. Since "October" was produced before "Enthusiasm", the soundtrack however much later, the presumption comes up that both films mutually influenced each other. This example illustrates that the proposed method is able to hint at correspondences between different films.

For the contemporary material the retrieved sequences contain fast and synchronously cut dialogue sequences (e.g. discussions and arguments between protagonists) and action sequences (fights, shootings, accidents). The extracted sequences are semantically important and may enrich further high-level tasks such as movie indexing, abstraction, and summarization.

VI. Conclusion

Directors employ the synchronous montage technique to increase the tension of a sequence and to highlight important events. The detection of such sequences enables a new way to extract semantically meaningful information from movies. We propose an approach for the automated extraction of such sequences based on a novel method for cross-modal temporal correlation estimation and a tolerant segmentation scheme. The retrieved sequences contain rich semantics which makes them suitable for high-level video abstraction.

In future, we will extend the evaluation to a larger set of movies and evaluate the benefit of the method to high level tasks such as movie summarization.

References

[1] D. Bordwell and K. Thompson, *Film art: an introduction*, 8th ed. McGraw-Hill, 2008.

[2] L. Fisher, "Enthusiasm: From kino-eye to radio-eye," in *Film Sound-Theory and Practice*, Weis and Belton, Eds. Columbia Univ. Press, 1985, pp. 247–264.

[3] Z. Barzelay and Y. Schechner, "Onsets coincidence for cross-modal analysis," *IEEE Trans. Multimedia*, vol. 12, no. 2, pp. 108 –120, 2010.

[4] G. Monaci and P. Vandergheynst, "Audiovisual gestalts," in *Proc. of the Int. Conf. on Computer Vision and Pattern Rec. Workshop*, 2006, p. 200.

[5] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," in *Advances in Neural Inf. Proc. Syst.*, 2000, pp. 813–819.

[6] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Adv. in Neural Information Proc. Syst.*, 2000, pp. 814–820.

[7] A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *Appl. Sig. Proc.*, pp. 1274–1288, 2002.

[8] P. Atrey, M. Kankanhalli, and R. Jain, "Information assimilation framework for event detection in multimedia surveillance systems," *Multimedia Systems*, vol. 12, pp. 239–253, 2006.

[9] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, pp. 345–379, 2010.

[10] C.-S. Perng, H. Wang, S. Zhang, and S. Parker, "Landmarks: A new model for similarity-based pattern querying in time series databases," in *Conf. on Data Engineering*, 2000, pp. 33–42.

[11] W. Fujisaki and S. Nishida, "Feature-based processing of audio-visual synchrony perception revealed by random pulse trains," *Visision Research*, vol. 47, no. 8, pp. 1075–1093, 2007.

[12] M. Zeppelzauer, D. Mitrović, and C. Breiteneder, "Analysis of historical artistic documentaries," in *International Workshop on Image Analysis for Multimedia Interactive Services*, 2008, pp. 201–206.

[13] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Int. Conf. on MM and Expo*, vol. 1, 2000, pp. 452–455.

[14] S. Eisenstein, *Film Form: Essays in Film Theory*. Harcourt Brace and Company, 1977, ch. A Dialectic Approach to Film Form, pp. 45–63.