Music-Guided Video Summarization using Quadratic Assignments

Thomas Mensink
University of Amsterdam
thomas.mensink@uva.nl

Thomas Jongstra
University of Amsterdam
jongstra@gmail.com

Pascal Mettes
University of Amsterdam
p.s.m.mettes@uva.nl

Cees G.M. Snoek
cgmsnoek@uva.nl
University of Amsterdam

ABSTRACT

This paper aims to automatically generate a summary of an unedited video, guided by an externally provided music-track. The tempo, energy and beats in the music determine the choices and cuts in the video summarization. To solve this challenging task, we model video summarization as a quadratic assignment problem. We assign frames to the summary, using rewards based on frame interestingness, plot coherency, audio-visual match, and cut properties. Experimentally we validate our approach on the SumMe dataset. The results show that our music guided summaries are more appealing, and even outperform the current state-of-the-art summarization methods when evaluated on the F1 measure of precision and recall.

CCS CONCEPTS

• **Information systems** → **Multimedia content creation**; • **Mathematics of computing** → Permutations and combinations;

KEYWORDS

Video Summarisation, Quadratic Assignment Problem

ACM Reference format:
Thomas Mensink, Thomas Jongstra, Pascal Mettes, and Cees G.M. Snoek. 2017. Music-Guided Video Summarization using Quadratic Assignments. In *Proceedings of ICMR '17, Bucharest, Romania, June 6-9, 2017,* 7 pages.
DOI: http://dx.doi.org/10.1145/3078971.3079024

1 INTRO

The goal of this paper is to create high-quality video summarizations, guided by an externally provided music-track. Consider for example that after a day of skiing with your GoPro camera, you reflect your mood by selecting a music-track and the computer will automatically create a video summary of your skiing day fitted on this specific music-track. Clearly a summary with classical music should have different dynamics, plots, and cuts than a summary based on funk music, even when the summaries are created from the same source video. Such an adaptive summarization method could be useful for creating video summarizations of social events by social media services or to personalize video summarizations.

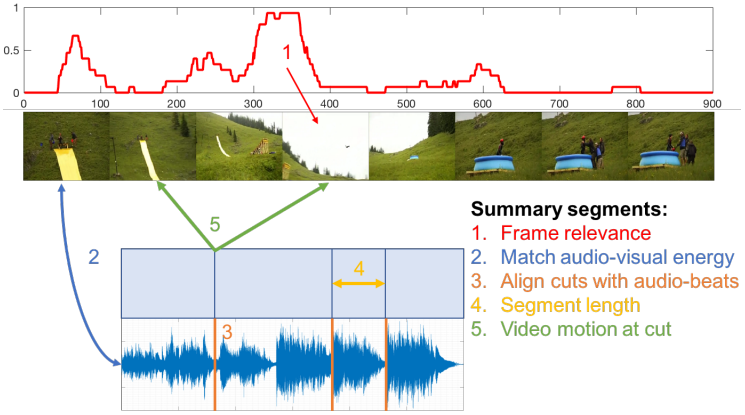

Figure 1: Illustration of key factors included in our model for automatically generating music-guided summaries.

The key key factors used in our music-guided summarization model are illustrated in Fig. 1. We are inspired by a large body of research focused on video summarization, either using only the visual source video [5, 8, 11–13, 18, 22], or combining multiple video modalities [4, 10, 20]. In contrast to these works, we aim to create a video summary fitted on a given music-track, which to the best of our knowledge has never been considered before.

1.1 Related work

Video summarization is often simplified to a frame interestingness problem [18], where interestingness can be measured using a variety of approaches, including object detections [11, 13], saliency [5, 8, 14], person detection [17], and landmark detection [12]. However, frame interestingness does not include any clues about the aesthetics of the summarization itself. Rather than relying exclusively on heterogeneous measures of interestingness for our video summarization, we emphasize on creating summaries, with a coherent plot and logical cuts.

Several works have previously incorporated coherency to make video summaries more viewer friendly. Such coherency can be performed by selecting sets of consecutive frames [5] or by adding temporal regularization [11]. The balance between interestingness and coherency, can be obtained using pre-segmentation methods [5], submodular optimization [6, 19] or recurrent neural networks [22].

ICMR '17, Bucharest, Romania

978-1-4503-4701-3/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078971.3079024

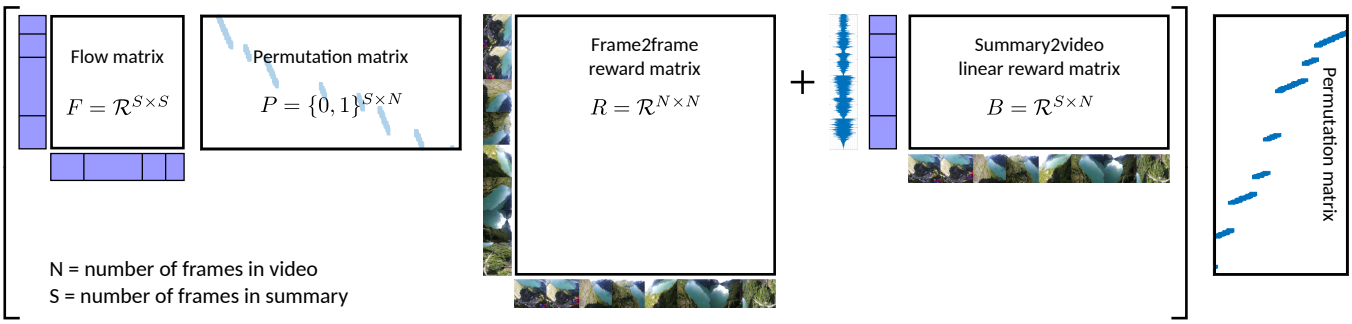

Figure 2: Overview of the Quadratic Assignment Problem formulation for audio-guided video summarization.

Here, we propose a model that jointly incorporates interestingness, coherency, and which is capable of adjusting the summaries based on a user-provided music-track.

Beyond exploiting just the visual content of videos for summarization, a number of works have proposed to jointly model the visual and audio modalities of videos [4, 10, 20]. For example, in [10] features from the visual, audio, and textual (subtitles) modalities are used to highlight interesting frames. In this work, we have an orthogonal goal, namely to align the summarization of videos with an externally provided music-track. Aligning video with a music-track is important for viewer experience, as indicated by [21].

Existing computational models for video summarization can not incorporate music-guidance. For example pre-segmentation methods are unlikely to yield segments which match the music properties, indicating we need a frame-based model. In contrast to seeing summarization as a selection process [5, 6], we see it as an assignment problem. Where for each summary slot the most suitable frame from the source video is assigned. Therefore we model our approach using the classical Quadratic Assignment Problem.

1.2 Contributions

We make the following contributions:

- we propose to summarize videos based on joint audio-visual information from the original video and a user-provided music-track,
- we model interestingness, coherency, and audio-match of the summaries jointly as a Quadratic Assignment Problem,
- we introduce measures for matching the dynamics and the beats of the music-track to the summary.

We include experimental evaluation to show that our approach is competitive to state-of-the-art summarization methods and generates audio-visual summaries that are tailored to the user provided music-tracks.

2 AUDIO-GUIDED SUMMARIZATION AS A QUADRATIC ASSIGNMENT PROBLEM

We see video summarization as classical assignment problem, where to each summary location $s \in S$, the most suitable frame $n \in N$ from the source-video is assigned. The suitability (or reward) of the assignment depends on three major factors:

- the interestingness of the frame;
- the match between the music in the summary and the visuals of the specific frame;
- the match with the previous frame in the summary, to model a story-line, segments, and cuts.

While frame-relevance yields a selection problem, and the music-video match can be modeled as a linear assignment problem, to model the cost for the subsequent frames we need to resort to quadratic assignment problems (QAP). Originally introduced for allocating facilities to certain locations in 1957 by Koopmans and Beckman [9], the QAP suits our model for music-guided video summarization. It enables to start from the summary, without the need of any pre-segmentation which would limit the flexibility to adjust to a specific music-track.

The QAP is a permutation problem, which aims to find the permutation $P \in \mathcal{P}$ with the highest reward:

$$f_{\text{QAP}}(P) = \text{tr}\left(\left(FPR^\top + B\right)P^\top\right), \tag{1}$$

Where, F denotes the flow matrix defined over the slots in the summary, and R defines a reward matrix between the frames in the source video, the last term B is a linear summary-source reward matrix, see Fig. 2 for an overview. Note that our permutation $P \in \mathcal{P}$ also encompasses a selection, since the summary contains less frames than the original source video. Therefore for any valid permutation P holds: $\sum_{j=1}^{N} p_{ij} = 1$ and $\sum_{i=1}^{S} p_{ij} = 1$.

Below, we introduce 6 summary components to model the three major factors of a good summarization, mentioned above. These components include, frame interestingness (I), Uniformity of storyline (U), Audio-Video Dynamics (AVD), Segment Length (SL), Motion Boundaries (MB), and Beat Cuts (BC). Each of these component is modeled as an assignment problem, and our final model is a weighted combination of these c components:

$$f_{\text{AVsummary}}(P) = \sum_c w^c f^c_{\text{QAP}}(P). \tag{2}$$

In general, solving a QAP, or its ϵ-approximate solution, is a NP-hard problem [16], in Section 2.4 we discuss our search strategy.

2.1 Frame interestingness

The first factor is the interestingness of each frame, for which we learn a frame-based interestingness classifier, using the human selected summaries as positive examples and the remaining frames as negative examples. For each frame we extract a set of M features,

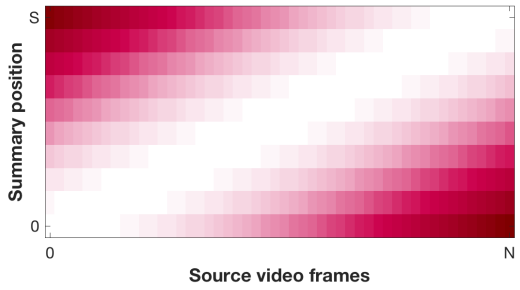

Figure 3: Illustration of the uniform reward B^{U}. Highest reward (white) obtained when following a uniform storyline.

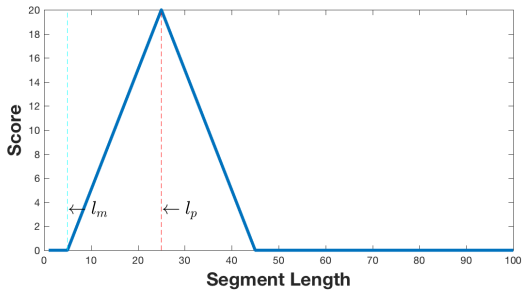

Figure 4: Illustration of *length prior*, given the segment length, the reward is based on l_m and l_p (5 and 25 in this fig).

and compute the interestingness score for frame k based on unary and pairwise terms, following [5, 11] we use:

$$i^k = w_0 + \sum_{m=1}^{M} w_m \, x_m^k + \sum_{m=1}^{M} \sum_{n=m+1}^{M} w_{mn} \, x_m^k \, x_n^k, \tag{3}$$

where x_m^k denotes the score of the m-th feature of the k-th frame.

Maximizing the interestingness of a summary is in principle (just) a selection problem, to remain within the QAP formulation, we define the linear matrix B^{I} to have identical values for each location s, given a frame k, as follows:

$$B^{\mathrm{I}}_{sk} = i^k. \tag{4}$$

Uniform Storyline. Besides the frame interestingness factor, we also observe that the source videos tell a roughly temporal uniform coherent story. This was also observed in [6], and indicates that a priori a good summary samples frames uniformly from the source video. Such a uniform sampling reward can be modeled as the linear part of the QAP, measuring the distance between summary position s and its uniform sampled frame $\tilde{p}_s$ from the source video:

$$\tilde{p}_s = \tfrac{N}{S}\left(s - \tfrac{1}{2}\right), \tag{5}$$

$$B^{\mathrm{U}}(s,n) = 1 - \tfrac{1}{N^2}\left(\tilde{p}_s - n\right)^2, \tag{6}$$

where B^{U} is normalized between 1 (when $n = \tilde{p}_s$) and 0, illustrated in Fig. 3. This could be seen as a prior model to retain uniform temporal coherency of the source video.

2.2 Music-Video Match

The second factor is the match between the music-track in the summary and the visuals of the source video. In this paper we aim to let the summary follow the audio dynamics, and therefore that the music dynamics should be similar to visual dynamics. To determine the audio dynamics we compute the relative amplitude for each summary frame location: $\hat{a}^s = a^s - \frac{1}{S}\sum_{s'} a^{s'}$, where a^s is the amplitude at time s. For the video dynamics we compute the relative motion per frame: $\hat{f}^n = f^n - \frac{1}{N}\sum_{k'} f^{k'}$, where f^n indicates the motion in frame n, based on the computation of the optical flow. We use the following linear rewards:

$$B^{\mathrm{AVD}}(s,m) = 1 - \gamma|\hat{a}^s - \hat{f}^m|, \tag{7}$$

where γ is a constant normalizing all rewards between 0 and 1.

2.3 Subsequent Frames

The last factor is to model the relation between two subsequent frames in the summary. This is important given that our model does not use pre-segmentations of the video. Subsequent frames could either form a consecutive segment, or define a cut (jump). The final model should balance between the length of series of consecutive frames, the placing of the cuts based on the frame features and based on the alignment with the music-track, and to represent the story-line of the original video.

Segment length. First, we model the length of a segment, since too short segments results in many cuts and makes the summary chaotic, while too long segments make the summary boring. Therefore we include a score based on the segment length l:

$$s(l) = \max(0, l_p - |l - l_p| - l_m) \tag{8}$$

where l is the length of a segment, l_p the prior length, and l_m the minimum length, see Fig. 4 (top). In QAP formulation, this entails the following flow and reward matrices:

$$R^{\mathrm{SL}}(m,n) = \begin{cases} 2 \text{ if } n = m + l_p, & 2 \text{ if } m = n + l_p, \\ 1 \text{ if } n = m + 2l_p - l_m, & 1 \text{ if } m = n + 2l_p - l_m, \\ 1 \text{ if } n = m + l_m, \text{and} & 0 \text{ otherwise,} \end{cases} \tag{9}$$

$$F^{\mathrm{SL}}(s,t) = \delta(s = t + l_m) - \delta(s = t + l_p) + \delta(s = t + 2l_p - l_m) - \delta(s = t - l_m), \tag{10}$$

where the flow matrix F uses the Dirac delta function $\delta(\cdot)$, which returns 1 if and only if the condition is true. The complexity of these matrices originate to ensure that coincidental rewards are canceled, *e.g.* a jump of exactly l_p frames would normally add an additional reward for a good sequence, yet due to the jump this needs to be zeroed.

Motion boundaries. Besides the segment length, we also have an aesthetic view on cuts, namely, a cut should take place when there is a minimum of motion in the frames, see Fig. 5. This is included in the QAP, by using the following flow and reward matrices:

$$R^{\mathrm{MB}}(m,n) = \frac{-1}{2}\delta(m > n+1)\,(x_m^m + x_n^m), \tag{11}$$

$$F^{\mathrm{MB}}(s,t) = \delta(s = t+1), \tag{12}$$

where x^m denotes the estimated motion magnitude, based on the KLT tracker. The reward matrix uses negative rewards between

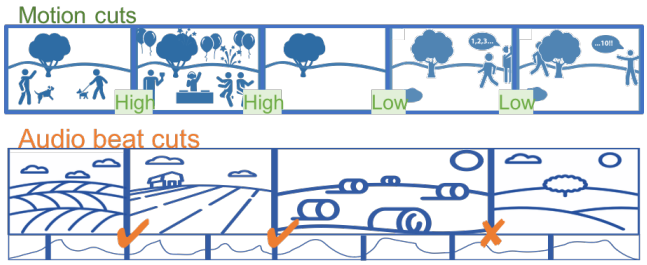

Figure 5: Illustration of *Motion boundaries* (top) and *Beat cuts* (bottom). Motion boundaries promote cuts between frames with low motion. Beat cuts promote cuts in the summary aligned with the music beat.

frames with high-motion. The flow matrix aggregates the motion magnitude scores for neighboring frames in the summary. This component is inspired by the pre-segmentation criterion used by [5].

Beat Cuts. A final consideration is the placement of cuts in the summary with respect to the music-track. We believe that cuts should be placed on the beat of the music, as illustrated in Fig. 5. Therefore we reward beat cuts with the following QAP:

$$R^{\text{BC}}(m,n) = \delta(m > n{+}1), \tag{13}$$

$$F^{\text{BC}}(s,t) = \delta(s = t{+}1)\ \delta(s \in \mathcal{B}), \tag{14}$$

where $\mathcal{B}$ denotes the set of summary slots matching a beat. The reward matrix returns a one, if and only if the frames in are not neighboring in the source video, and F aggregates these scores only when there is a beat.

2.4 Greedy search

The search space, for a summary of S frames, from a source video with N frames, is huge, even though we constraint the search space to a forward selection process only. Consider a 5 minute source video (N=9000) and a 30 seconds summary (S=900), for the first frame we can select (N-S) frames, the second frame is then selected from (N-f_1-S-1), *i.e.* pick any frame between the previous selected frame and the S-1 last frames, etc etc. This is in the order of $O(N!/(N-S)!)$, and yield zillions of possible permutations. As said before, for a general QAP (or its ϵ-approximate solution) there exist no efficiently optimal solution. Therefore we resort to a greedy search which selects one frame at the time.

Given a partial permutation P_g, consisting of the permutation of the first g summary frames, we add the next frame $p_{g+1} = t$. We select the best frame t, based on the current reward and an approximation of the (expected) future rewards. We approximate the future rewards of frame t, by evaluating Eq. (2) with the partial permutation when adding frames t, t+1, t+2, ... consecutively to the summary, to the manifestation of the next 1 to 4 beats. This yields per frame t four scores, which we normalize for their different lengths, and we select the frame with the highest score. We call our approximation the *beat-look-ahead* score.

To further reduce the search we (1) require each segment to last at least one second, *i.e.* for a new segment the next 25 frames are also added, this is identical to adding a reward of $-\infty$ on segments shorter than $l_m = 25$; and (2) consider only frames in a small window around the previous selected frame, if $p_g = v$, then frame t is selected from $v + 1 \le t \le v + \frac{N}{5}$. This window size is based on preliminary experiments.

Figure 6: A frame from each of the 25 videos of SumMe [5].

3 EXPERIMENTS

In this section we experimentally validate our proposed methods. First, we describe the experimental setup. Then, we present *visual-only* summarization results to compare to recent work. Finally, we present the results of our music-guided video summarization.

3.1 Dataset and experimental setup

SumMe dataset [5]. For all experiments we use the recently introduced SumMe dataset, which contains 25 amateur-shot raw videos (40-400 seconds) covering holiday moments, events and sports. The videos are illustrated in Fig. 6. For each video, 15-18 human participants have been asked to create a video summary of 5-15% of the original length. The diversity of the videos and the availability of multiple annotations make the dataset perfectly suited for illustrating performance of our summarization methods.

Evaluation. To evaluate a video summarization we use the provided human summaries. First, we compute precision and recall of frames between a generated summary and a human summary, and use the F1-score to balance precision and recall. To incorporate the fact that summaries are highly subjective, we follow [6] and use the highest F1 score between an automatically generated summary and any of the human summaries, we denote this as the **Best F1** score. It ensures that the generated summary is rewarded if it matches closely to one of the human annotators.

	Best F1	Recall	Precision
Random	16.8	17.5	16.6
Uniform	27.1	29.4	25.1
Gygli et. al. [5]	39.3	44.4	35.3
Gygli et. al. [6]	39.7	43.0	36.8
Single Frames*	34.7	38.4	33.5
Super Frames* [5]	36.4	40.4	34.2
QAP Model*	38.3	42.4	36.6

Table 1: Comparison to state-of-the-art. Methods indicated with * use the same features. Our QAP model allows to directly use frame-based interestingness prediction, without resorting to pre-segmentation methods

Interestingness Features. For the interestingness prediction, we use a subset of the features used in [5]. This is a collection of features modeling *attention*, using spatial and temporal salience [3, 7]; and modelling *aesthetics*, based on colourfulness, contrast, and distribution of edges [1, 8]. The other features used in [5], *a.o.* landmark detection and person detection, were not reproducible, and therefore not used in our experiments.

We extend this collection of features, with a high-level frame description based on ImageNet objects [2]. In order to do so, we extract per frame the penultimate layer of the deep network, provided by [15]. This is a 1024 dimensional feature, which we reduce to 64 dimensions using PCA, so that we can learn both the unary and pairwise terms in the interestingness prediction.

We learn the parameters w of Eq. (3), by random sampling 100 frames from each video in the dataset and train a predictor. The final model is an average over 50 repetitions of this training.

3.2 Visual-only Summaries

In this set of experiments we evaluate the performance of our proposed models and compare them to the current state-of-the-art.

Tuning summarization components. In a set of preliminary results, we tune the weights, used in Eq. (2), of each component. Starting with an equal weighting ($\forall c: \ w_c = 1$), we tune the components one by one. For each component, we use leave-one-out performance to vote for a specific parameter value, and the value with the highest number of votes is selected. Since parameters interact, repeating this search could result in different weight values. The obtained weights are: I = 1, MB = 2, SL = 1, and U = $\frac{1}{2}$.

Experimental results. The goal of this experiment is to see whether the QAP model is able to generate high quality summarizations without the need for pre-segmentation methods. We compare our QAP model with current state-of-the-art methods on this dataset and we add a model based on *Single Frame* interestingness predictions and on the *Super Frames* pre-segmentation used in [5]. The latter methods use exactly the same raw interestingness predictions, which makes them comparable.

The results are presented in Table 1. We observe that any of the computational methods outperform random or uniform segmentation. Furthermore we observe that our implementation of [5] (denoted as Super Frames) scores about 3% worse compared to the published results, this is due to differences in the used features. Finally, when looking at models using the same features, we observe that adding structure to the single frame interestingness predictions is beneficial, given that super frames outperform single frames, and our QAP model outperforms both.

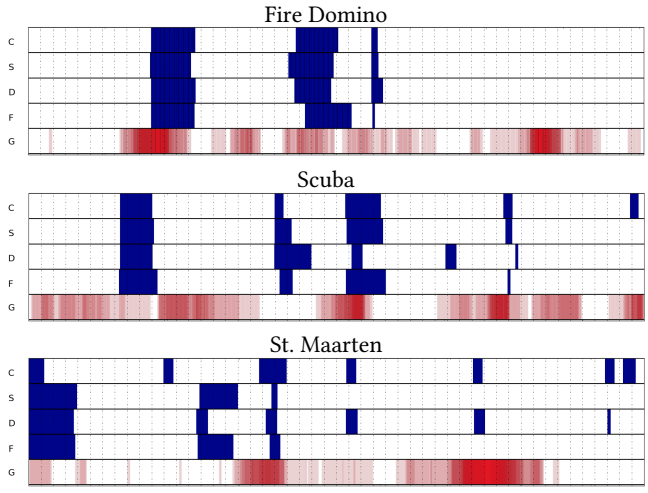

Figure 7: Illustration of summaries guided by different music styles. Selected frames are indicated in blue, the (average) human summary in red. Clearly our model combines frame interestingness with music properties (energy / beats), resulting in different summaries for different songs.

In Table 2, we show per video performance of our different models, *Single*, *Super Frames*, and *QAP*. We observe that in general, the performance of the QAP model is superior. However, we can see some outliers: the single model performs exceptionally well on the video "Playing ball" (75.8% vs 36.0% for QAP), the super frames on the video "Eiffel tower" (47.9% vs 30.6% for QAP), and the QAP model for the "Jumps" video (93.8% vs 41.0/35.7% for single/super frames). These large differences indicate that the relative importance of specific model components vary for some of these videos.

3.3 Music-Guided Summaries

In this set of experiments we validate our model for creating summaries guided by an externally provided soundtrack, we coin our method **QAP-AV**. For illustrating the influence of different styles, we use four songs of different genres: classical, swing, funk and downtempo. In the provided *supplementary material*, we show a composition of several summaries fitted on these music-tracks.

Weighing audio-visual summarization components. To determine the weights of the different components we resort to the same heuristic grid search as used for the visual-only method. We start from an equal weighting ($\forall c: \ w_c = 1$) of all components, and tune all weights based on the *swing* music-track, using leave-one-out voting based on the **Best F1** score. The obtained weights are: I = 5, MB = 2, SL = 5, U = 1, AVD= 5 and BC= 1.

Results. In Table 2, we show the results of the **QAP-AV** method evaluated for the four different music-tracks for each video. For comparison we have also included the visual-only QAP models.

		Visual only			Music - Guided				
		Single Frames	Super Frames [5]	QAP	Funk	Downtempo	Swing	Classical	*Best Song*
ego.	Base jumping	21.7	32.7	21.6	**31.0**	28.9	29.3	29.0	***31.0***
	Bike polo	37.5	39.2	39.0	**47.9**	47.7	45.7	46.3	***47.9***
	Scuba	39.3	56.5	37.2	**57.0**	44.4	55.7	56.7	***57.0***
	Valparaiso downhill	26.3	29.8	26.9	39.7	**43.1**	36.1	33.3	***43.1***
moving	Bearpark climbing	27.6	29.3	24.3	**33.0**	26.9	22.3	28.4	***33.0***
	Bus in rock tunnel	39.5	**46.1**	34.9	37.8	43.3	39.4	42.5	*43.3*
	Car rail crossing	**39.3**	30.6	31.0	22.0	23.4	20.0	23.8	*23.8*
	Cockpit landing	38.8	39.9	52.0	53.5	56.0	56.4	**56.8**	***56.8***
	Cooking	20.9	22.3	30.2	27.7	27.2	**36.6**	25.6	***36.6***
	Eiffel tower	42.3	**47.9**	30.6	42.1	40.8	42.4	43.9	*43.9*
	Excavators river crossing	22.6	23.3	**33.5**	29.0	28.9	28.9	28.8	*29.0*
	Jumps	41.0	35.7	**93.8**	70.8	71.5	71.5	72.9	*72.9*
	Kids playing in leaves	38.2	39.3	28.4	47.3	**50.3**	**50.3**	48.4	***50.3***
	Playing on water slide	44.0	44.0	35.3	24.2	48.3	**50.1**	**50.1**	***50.1***
	Saving dolphins	21.3	29.3	**46.9**	23.6	25.2	27.5	29.8	*29.8*
	St. Maarten Landing	37.9	26.7	39.6	28.1	**44.6**	26.2	42.3	***44.6***
	Statue of Liberty	22.7	**26.0**	24.4	25.0	22.2	22.6	16.7	*25.0*
	Uncut evening flight	19.3	20.4	**36.8**	33.3	33.8	34.6	33.3	*34.6*
	Paluma jump	25.0	21.9	**50.6**	36.3	35.8	37.6	30.5	*37.6*
	Playing ball	**75.8**	72.0	36.0	56.2	66.2	65.4	67.5	*67.5*
	Notre Dame	**28.6**	28.2	21.8	23.6	22.1	22.4	22.9	*23.6*
static	Air Force One	32.1	**36.2**	27.3	35.0	26.4	24.3	24.4	*35.0*
	Fire domino	40.8	44.2	49.7	62.2	67.2	69.7	**71.0**	***71.0***
	Car over camera	44.6	48.7	65.8	66.3	68.1	67.3	**70.6**	***70.6***
	Paintball	36.2	40.1	39.3	**48.9**	40.5	42.7	41.6	***48.9***
	Mean Best-F1	34.7	36.4	38.3	40.1	41.3	41.0	**41.5**	***44.3***
	Segment statistics								
	Avg. # of cuts	154.4	5.1	8.6	5.9	6.5	6.5	6.8	
	On-beat cuts (%)			± 5*	89.8	94.4	99.4	95.9	

Table 2: Per video results on the SumMe dataset using the Best F1 measure, comparing *visual-only* to *audio-guided* methods on 4 different audio-tracks. (*) Average on-beat cuts over all songs (for QAP). Surprisingly the audio-guided methods outperform the visual-only methods by up to 6% absolute performance when an oracle could provide the most suited audio-track per video.

First, we observe that our music-guided methods, surprisingly, obtain better results than any of visual-only methods, even better than the state-of-the-art summarization method of [6] (39.7%, see Table 1). Our results are also better than the recent work of [22], where 38.6% is reported using slightly different evaluation settings and on par with their method that uses extensive additional labeled data to train interestingness and hyper-parameters (41.8%). Further, the relatively stable average **Best F1** for the different music-tracks, indicate that the weighting of the components is not music-track specific per se. Still the large variation in performance between different models for a specific video, indicate that the relative importance of model components vary. Finally, when we compare statistics about the cuts in the video summary, it is apparent that the no-music QAP model generates more cuts, and have hardly any aligned with the audio (averaged over all songs), while the music guided model has almost all cuts aligned with music. In conclusion, the music guidance (in audio-visual match and beat cuts), enables to generate higher quality video summarizations.

In Fig. 7 we show the selection of frames for the QAP-AV model for different soundtracks for three different videos, and as reference show the average human summary selection. Video examples are included in the supplementary material.

4 CONCLUSION

In this paper we have introduced a model for music-guided video summarization, which we have modeled as a quadratic assignment problem (QAP). The QAP formulation allows to dynamically create video segments, match the music-dynamics, have boundaries with low motion, and with cuts (mostly) on the beat of a provided music-track. Experimentally we have validated our approach on the SumMe dataset, showing that our QAP model is on par with current state-of-the-art video summarization and that our music-guided models even outperform these. In conclusion our QAP model yields high quality summaries (in terms of F1), which are also more appealing to watch (see examples in supplementary material).

For future work, we aim to extend our method to exploit the audio track of the source video as input modality, to allow for repetitions (*e.g.* for the chorus of a song), and to include basic video effects (panning, crop, zoom, speed-up and slowmotion). This will search the limits of the QAP, since rewards for these effects require higher-order dependencies than only neighboring dependencies, which could be based either on parametrized flow and reward matrices, or by another search strategy to evaluate permutations.

REFERENCES

[1] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2006. Studying Aesthetics in Photographic Images Using a Computational Approach. In *ECCV*.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

[3] N. Ejaz, I. Mehmood, and S. W. Baik. 2013. Efficient visual attention based framework for extracting key frames from videos. *Signal Processing: Image Communication* 28, 1 (2013), 34–44.

[4] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. 2013. Multimodal Saliency and Fusion for Movie Summarization Based on Aural, Visual, and Textual Attention. *IEEE Trans. Multimedia* 15, 7 (2013), 1553–1568.

[5] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. 2014. Creating summaries from user videos. In *ECCV*. Springer, 505–520.

[6] M. Gygli, H. Grabner, and L. Van Gool. 2015. Video summarization by learning submodular mixtures of objectives. In *CVPR*. 3090–3098.

[7] X. Hou, J. Harel, and C. Koch. 2012. Image signature: Highlighting sparse salient regions. *IEEE Trans. PAMI* 34, 1 (2012), 194–201.

[8] Y. Ke, X. Tang, and F. Jing. 2006. The Design of High-Level Features for Photo Quality Assessment. In *CVPR*.

[9] T. Koopmans and M. Beckmann. 1957. Assignment problems and the location of economic activities. *Econometrica* (1957), 53–76.

[10] P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos, and A. Potamianos. 2015. Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization. In *ICIP*. IEEE, 4361–4365.

[11] Y. Lee, J. Ghosh, and K. Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *CVPR*. 1346–1353.

[12] F. Liu, Y. Niu, and M. Gleicher. 2009. Using Web Photos for Measuring Video Frame Interestingness.. In *IJCAI*.

[13] Z. Lu and K. Grauman. 2013. Story-driven summarization for egocentric video. In *CVPR*. 2714–2721.

[14] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. 2002. A user attention model for video summarization. In *MM*.

[15] P. Mettes, D. Koelma, and C. Snoek. 2016. The ImageNet Shuffle: Reorganized Pre-training for Video Event Detection. In *ICMR*.

[16] S. Sahni and T. Gonzalez. 1976. P-complete approximation problems. *Journal of the ACM (JACM)* 23, 3 (1976), 555–565.

[17] M. Smith and T. Kanade. 1998. Video skimming and characterization through the combination of image and language understanding. In *Content-Based Access of Image and Video Database*.

[18] B. Truong and S. Venkatesh. 2007. Video abstraction: A systematic review and classification. *ACM TOMCAPP* (2007).

[19] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M Rehg, and Vikas Singh. 2015. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *CVPR*.

[20] Li Y. and Merialdo B. 2012. Video Summarization Based on Balanced AV-MMR. In *International Conference on Multimedia Modeling (MMM)*.

[21] J. You, U. Reiter, M Hannuksela, M. Gabbouj, and A. Perkis. 2010. Perceptual-based quality assessment for audio–visual services: A survey. *Signal Processing: Image Communication* (2010).

[22] K. Zhang, W.-L Chao, F. Sha, and K. Grauman. 2016. Video Summarization with Long Short-Term Memory. In *ECCV*.