# The effect of different video summarization models on the quality of video recommendation based on low-level visual features

Yashar Deldjoo
Politecnico di Milano
Via Ponzio 34/5
Milan 20133, Italy
yashar.deldjoo@polimi.it

Paolo Cremonesi
Politecnico di Milano
Via Ponzio 34/5
Milan 20133, Italy
paolo.cremonesi@polimi.it

Markus Schedl
Johannes Kepler University
Altenberger Street 69
Linz 4040, Austria
markus.schedl@jku.at

Massimo Quadrana
Politecnico di Milano
Via Ponzio 34/5
Milan 20133, Italy
massimo.quadrana@polimi.it

## ABSTRACT

Video summarization is a powerful tool for video understanding and browsing and is considered as an enabler for many video analysis tasks. While the effect of video summarization models has been largely studied in video retrieval and indexing applications over the last decade, its impact has not been well investigated in content-based video recommendation systems (RSs) based on low-level visual features, where the goal is to recommend items/videos to users based on visual content of videos. This work reveals specific problems related to video summarization and their impact on video recommendation. We present preliminary results of an analysis involving applying different video summarization models for the problem of video recommendation on a real-world RS dataset (MovieLens-10M) and show how temporal feature aggregation and video segmentation granularity can significantly influence/improve the quality of recommendation.

## CCS CONCEPTS

• **Information systems → Recommender systems**;

## KEYWORDS

Content-based video recommendation, temporal feature summarization, feature aggregation, shot segmentation granularity, evaluation

**ACM Reference format:**
Yashar Deldjoo, Paolo Cremonesi, Markus Schedl, and Massimo Quadrana. 2017. The effect of different video summarization models on the quality of video recommendation based on low-level visual features. In *Proceedings of CBMI, Florence, Italy, June 19-21, 2017,* 6 pages.


*CBMI, June 19-21, 2017, Florence, Italy*

ACM ISBN 978-1-4503-5333-5/17/06…$15.00
https://doi.org/10.1145/3095713.3095734

https://doi.org/10.1145/3095713.3095734

## 1 INTRODUCTION AND CONTEXT

A video can be represented by a set of frame-level (visual) features $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{d*n}$ where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the $i$-th frame with $d$-dimensional feature description and $n$ is the total number of frames in the video. For a regular video, $n$ is usually in the order of tens of thousands where consecutive frames share a lot of similarity in their visual characteristics, unless at transitioning frames from one shot to another shot. Techniques have been sought to organize video features into more compact forms as a first step for a number of tasks, including video retrieval and video classification [7, 8, 11, 12].

In this paper we focus on the problem of video summarization for the video recommendation domain [3, 18]. Content-based video recommender systems (RS) are a class of information filtering techniques that provide personalized video recommendations to users by building a profile of the user from the content-based descriptions of items/videos. Machine learning techniques are used to learn the user profile from both (i) the videos the user has previously interacted with (*e.g.*, liked, bought *etc.*) and (ii) the features of each video. The user profile is a vector in the *feature space* and shows user's interests to each of the feature components.

We argue that in building video recommendation algorithms based on low-level visual features, quality of recommendations depends on how video features are represented across the video stream. The procedure typically adopted to extract visual features can be summarized into three steps:

- Shot boundary detection is used to segment a video temporally and it is intrinsically linked to the way the video is produced. Its goal is to segment a video into meaningful parts, and thus it is very often the first step in algorithms that accomplish other video analysis tasks. A standard technique to detect shot boundaries is to calculate the similarity between every two consecutive frames. When the similarity goes below a certain threshold, a shot change is detected. A high similarity threshold would result in a fine-grained

segmentation, whereas a low threshold would yield a coarse segmentation. For example, for a random video in our dataset (see Section 2 and 3) a (histogram) similarity threshold of $t = 0.45$ results in 96 shots whereas a similarity threshold of $t = 0.95$ produces 1936 shots. The amount of information transferred into a recommender system can be largely different by choosing a fine-grained or coarse segmentation.
- After a video is segmented into meaningful shots, each shot is represented by a feature descriptor calculated from one or more frames within the shot . In the former case, the selected frame is regarded as the key frame, whereas in the latter case, the function that combines the frames or features is known as an *aggregation* function.
- After each shot has been represented with a descriptor of fixed length, the shots' descriptors can be aggregated by using different aggregation functions. The key difference between this stage and the previous stage is that features across shots vary significantly over time, whereas the features within a shot do not change significantly. Therefore, the aggregation function that combines features across shots needs to be of different nature and should capture the underlying statistics of the features across the video stream.

Based on the discussion above, we articulate the following high-level research hypothesis: *We can improve the quality of recommendation with respect to MSE in content-based video recommender system based on visual features by using a summarized model of the video that takes into account the importance of between-shot and within-shot feature aggregation and shot segmentation granularity*. We articulate the research hypothesis along three research questions:

**RQ1: Sensitivity to Shot-Segmentation Granularity**. We segment all videos in our movie dataset into shots in a *coarse-to-fine* fashion to determine an optimal shot segmentation level that can improve the quality of recommendation.

**RQ2: Effect of Shot Representation**. As there exists considerable amount of temporal redundancies between shot frames, we investigate if features extracted from a single key frame of each shot result in the same quality of recommendation than averaging the features extracted from all the frames within the shot.

**RQ3: Effect of Video Representation**: At this stage, a number of important questions are raised. First, the number of video shots produced after video segmentation is different on different videos depending on the video length, how the video is filmed (*e.g.* if it contains many or few camera movements) and how it is segmented (coarse or fine). The key question is how to combine these shot-level feature vectors which are different in number in different videos into a *fixed length* descriptor to represent each video? Second, is it sufficient to select *a fixed number* of shots in a random fashion and use the information from those shots as representation of the entire video? Third, as the variation of features across shots is significantly higher than within shots, which is the best aggregation strategy across shots?

This paper does not assert to be proposing an entirely novel method as the other works in video and music indexing and retrieval have tried to, e.g. [8, 12, 14, 16]. However, we think it will be instructive in the future to explore many video summarization schemes presented in this paper in the context of video recommendation based on visual features. In fact, to date the authors know about no other video recommender system based on visual features which has exploited the effect of these summarization models in the recommendation context. The contributions of this paper are two-fold:

- We articulate specific problems related to video recommender systems based on visual features currently deployed.
- We present preliminary results that show how temporal feature aggregation and video segmentation level can significantly influence/improve the quality of recommendation.

## 2 METHODOLOGY

The methodology adopted to evaluate the impact of using different video summarization models on the quality of video recommendation comprises of *five* steps:

1. Video Shot Segmentation
2. Visual Feature Extraction
3. Within-Shot Feature Aggregation
4. Between-Shot Feature aggregation
5. Recommendation

We perform feature aggregation in two levels: *within* and *between* shots, where in the former the effort is to reduce the redundancy that exist between frames of a shot (due to high correlation between successive frames) and in the latter the attempt is to take advantage of the diversity of feature values across shots (due to independence of frames across shots).

We have evaluated the quality of recommendation with respect to following: (i) Shot-Segmentation Level, (ii) Within-Shot Feature Aggregation Type and (iii) Between-Shot Feature Aggregation Type. The flowchart of the methodology is shown in Figure 1.

### 2.1 Video shot segmentation

The initial step toward the goals presented in Section 2 and enable extraction of features is to segment video streams into shots. A great number of methods have been proposed in the past years [2, 10]. The color histogram distance is one of the most reliable variants used as a measure of (dis)similarity between consecutive video frames for the purpose of content-based video retrieval, object recognition, and others. The basic idea is that the video content intensity does not change rapidly within but across shots. Thus, hard cuts and other short-lasting transitions are detectable as a signal peak in the time series of the differences between intensity histograms of two consecutive frames. A histogram is computed for each frame in the video and the *histogram intersection* is used as the means of comparison (measure of local activity) according to Equation 1,

$$\mathbf{s}(h_t, h_{t+1}) = \sum_b \min(h_t(b), h_{t+1}(b)) \tag{1}$$

where $h_t$ and $h_{t+1}$ are histograms of successive frames and $b$ is the index of the histogram bin. By comparing $\mathbf{s}$ with a predefined threshold and letting this threshold vary from a small value (0.45) to a large value (0.95), we segment the videos in our video dataset into shots from a coarse to fine fashion.

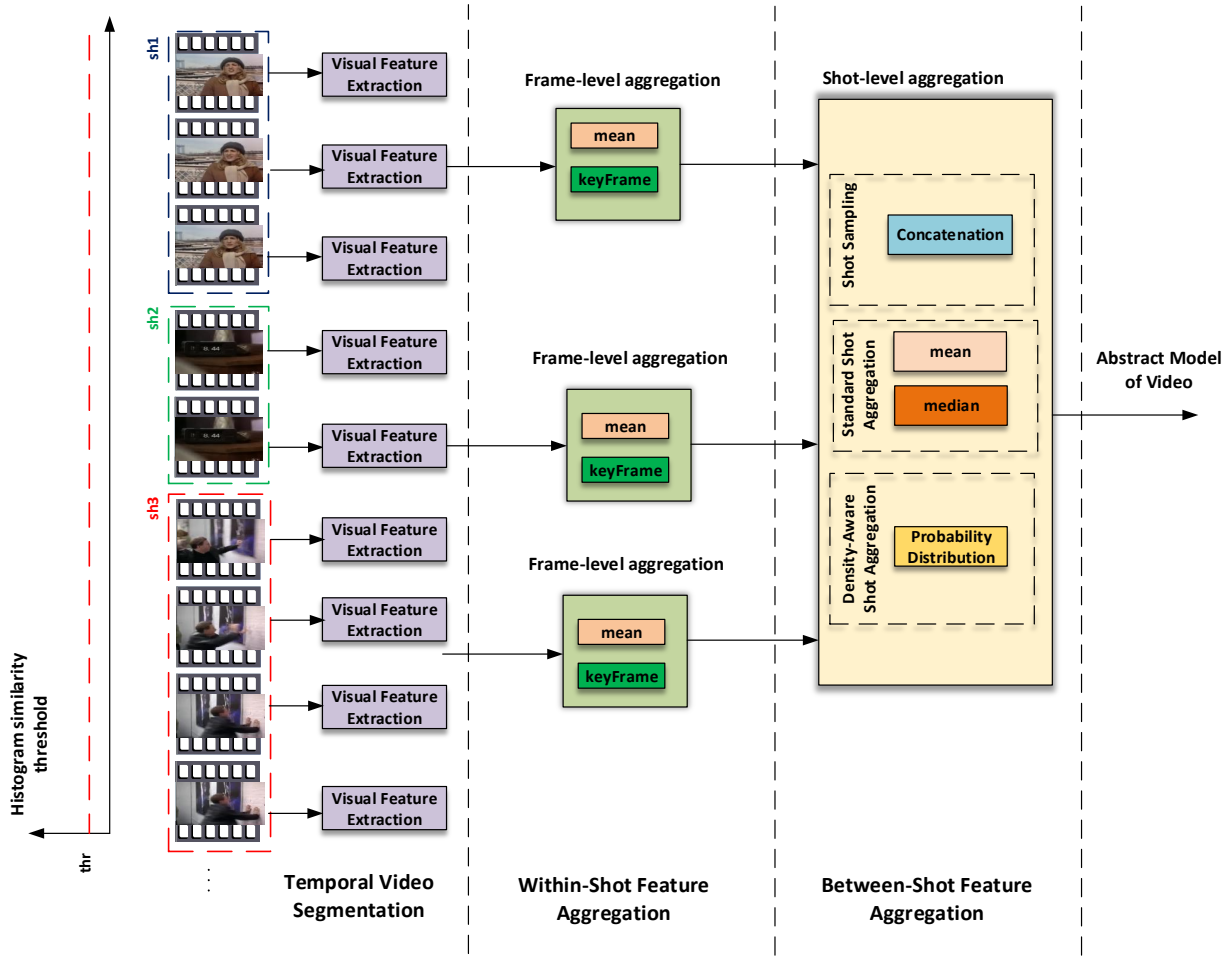


**Figure 1: Elements of the video summarization model in our experiment. A video is first segmented into shots by using a coarse-to fine segmentation granularity level. From each frame in one shot, a number of predefined visual features are extracted. The frame-level visual features are aggregated by using a within-shot aggregation method. Finally, the shot level features are aggregated into a fixed-length descriptor by using different between-shot aggregation strategies.**

## 2.2 Visual features extraction

In order to optimize the contribution of visual features to video recommendation, two categories of visual features are investigated: *static features* and *dynamic features*.

- The *static dimension* includes cues that are salient because of a change over *image space*; for example, color variation of all pixels in an image [1].
- The *dynamic dimension* includes features that are salient because of change over *time*.

In our experiments, we have selected four categories of low-level visual features quantified in six feature variations as described in Table 1. Two of these four feature categories capture the static aspect of videos, whereas the other two reflect the dynamic aspect [5, 10]. As for the motion feature, we use the standard optical flow technique to estimate motion vectors in each video frame. The motion feature is represented by calculating the *mean*, *median*, and *standard deviation* of motion vectors. Regarding shot duration, we

**Table 1: Visual features used in our study.**

| | feature | aggregation within frame | type |
|---|---|---|---|
| 1 | color variance | - | static |
| 2 | lighting key | - | static |
| 3 | motion | mean | dynamic |
| 4 | | median | dynamic |
| 5 | | std | dynamic |
| 6 | shot duration | - | dynamic |

measure the length of each shot (in terms of number of frames) as a characteristic of that shot (different from previous works which measure the average shot length on the entire video). Finally, all the features are normalized to the range [0-1] by using the min-max normalization scheme. Previous works have shown the effectiveness of these low-level visual features in recommendation and classification contexts [3, 4, 15, 17].

## 2.3 Within-shot feature aggregation

We select and compare the functioning of two within-shot aggregation functions: *mean* and *key frame selection*, where a key frame is chosen as the middle frame of a shot [19]. The reason for this choice is to investigate if it is sufficient to use a key frame for shot representation. This is a common approach in video indexing and retrieval applications and here we are comparing the performance in the context of recommendation.

## 2.4 Between-shot feature aggregation

In our experiments, we investigate three strategies:

- *Shot sampling*. With shot sampling, a predefined number of shots are randomly selected and their corresponding feature vectors are concatenated to create a *super vector* of raw features.
- *Averaging aggregation*. All feature vectors are aggregated using either mean or median.
- *Density-aware feature aggregation*. In this approach, we estimate a probability density function (PdF) with $B$ bins for each of the visual features across shots. For this, we map each of the six visual features in each shot, to one of $B$ bins of a PdF. We repeat this step for all shots until the complete PdF is built for that particular feature. This results in six PdFs for six features used in our experiment. Then we concatenate the corresponding PdFs for each feature to create a super vector. The advantage of the PdF-approach is three-fold: (1) it considers all video shots, (2) it maps videos with different number of shots into a fixed length feature vector representation, (3) it captures the underlying statistics of the features across shots. We chose the number of histogram bins in our experiment 8, and 16 denoted by PdF 8 and PdF 16.

## 2.5 Recommendation Model

The recommendation score of an unrated item $i$ for user $u$ is computed as a linear model as shown in Equation 2,

$$\widetilde{\mathbf{r}}_{ui} = \mu + \mathbf{b}_i + \mathbf{b}_u + \mathbf{p}_u \mathbf{f}_i \tag{2}$$

where $\widetilde{\mathbf{r}}_{ui}$ is the estimated rating for user $u$ on video $i$, $\mathbf{f}_i \in \mathbb{R}^{n_f}$ is the feature vector of item $i$ in which $n_f$ is the number of features and $\mathbf{p}_u \in \mathbb{R}^{n_f}$ is the *profile of user u*, a weight vector which measures the user $u$'s taste on each of the feature vector components. The user profile $\mathbf{p}_u$ is estimated by the ridge regression optimization model [13], shown in Equation 3,

$$\underset{\mathbf{p}_u}{\text{minimize}} \ \frac{1}{2}\|\mathbf{r}_{ui} - \mathbf{p}_u \mathbf{f}_i\|_2^2 + \lambda \|\mathbf{p}_u\|_2^2 \tag{3}$$

where $\|.\|_2^2$ is the $\ell_2$-norm and the constant $\lambda > 0$ is the regularization parameter. The user profile $\widetilde{\mathbf{p}}_u$ is learnt in the training phase and used to predict unknown rating in the test phase.

**Table 2: Characteristics of the dataset used in the evaluation. rsize and csize are the average number of ratings per user and item respectively.**

| dataset | #users | #items | #ratings | rsize | csize | density |
|---|---|---|---|---|---|---|
| **ML-10M** | 27992 | 1459 | 281734 | 10.06 | 193.10 | 0.68 % |

# 3 EXPERIMENTAL RESULTS

We evaluate the performance of the recommender system on the MovieLens-10M dataset [9], which contains user–item interactions of users with an up-and-running movie recommender system. For each movie in the dataset, we have extracted the video features from the corresponding trailer downloaded from YouTube, if available. The characteristics of the final dataset is shown in Table 2.

## 3.1 Evaluation Methodology and Metrics

We employ 10-fold cross validation (CV) in our experiments. Ratings are therefore partitioned into 10 non-overlapping subsets. In each run, 90% of the ratings are included in the training set and 10% of the ratings are used for testing. The evaluation is conducted by comparing the Mean Squared Error (MSE) between the true rating $\mathbf{r}_{ui}$ and the predicted rating $\widetilde{\mathbf{r}}_{ui}$

$$MSE = \frac{\sum(\widetilde{\mathbf{r}}_{ui} - \mathbf{r}_{ui})^2}{n} \tag{4}$$

where $n$ is the number of ratings in the test set.

**Table 3: Comparison of the effect of within- (left to right) and between-shot (top to bottom) aggregation models on the quality of video recommendation. The results reported are MSE values based on 10-fold CV experiments.**

| Video Summarization Models | | MSE | |
|---|---|---|---|
| | | within-shot aggregation | |
| | | key frame (1 frame) | mean (multiple frames) |
| between-shot aggregation | mean | 0.797 | 0.796 |
| | median | 0.797 | 0.793 |
| | shot sampling 8 | 0.806 | 0.813 |
| | shot sampling 16 | 0.771 | 0.772 |
| | PdF 8 | 0.770 | 0.771 |
| | PdF 16 | **0.756** | **0.755** |

## 3.2 Comparison of the Effect of Within- and Between-Shot Aggregation Models

The comparison of the effect of different video summarization approaches as a function of within- and between-shot aggregation models is shown in Table 3. Shot sampling 8 and 16 mean, respectively, 8 and 16 shots were sampled randomly from the videos under consideration in order to aggregate the shot feature vectors. PdF 8 and 16 refer the number of bins (8 or 16 bins) used in generation of feature histograms. We summarize the results based on Table 3 as follows:

1. *Within-shot Aggregation Model*:
   - **Best Model**: It can be seen that the *key frame* and *mean* within-shot aggregation models perform almost similarly on all cases with negligible differences and it is hard to prefer one over the other.
   - **Explanation**: The result above can be explained by the fact the frames within shots contain a high-level of similarity in their visual appearance.
   - **Assumption**: From this point on, we will report all the results based only on *key frame* within-shot aggregation model.
2. *Between-shot Aggregation Model*:
   - **Best Model**: We performed a multiple comparison test based on the results reported in Table 3 with the goal of understanding best within-shot aggregation model. We used 1-way ANOVA to investigate whether there is a significant difference between the means of six between-shot aggregations models. Results are summarized in Figure 2, where we can see that the *PdF16* between-shot aggregation approach outperforms other approaches with statistically significant difference ($p < 0.01$).

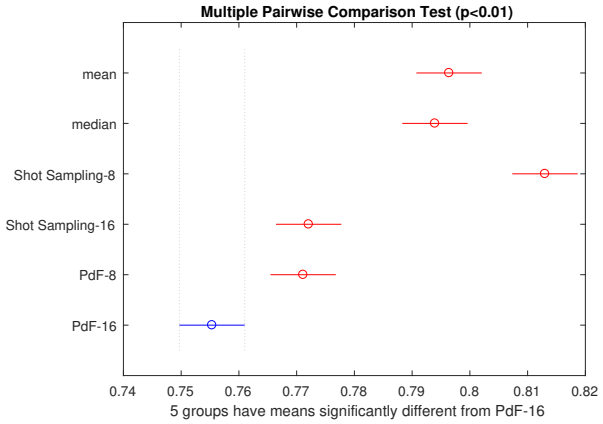


**Figure 2: Results of ANOVA test for the between-shot aggregation models.**

   - **The effect of descriptor length**: We also report the lengths of the final video descriptors in Table 4. By comparing the results presented above and the information provided in Table 4, the following conclusions can be drawn:
     - The summarization approach which builds a probability density function (PdF) of the feature values *across* shots, is the optimal choice as the between-shot aggregation function. For instance, we can see that *PdF8* and *shot sampling 16* offer equal performance, but the former's video descriptor length is only half of the latter's.
     - Our study does not support the notion that the larger the (final) video descriptor, necessarily it would be the better the quality of recommendation. In fact, in some cases (*e.g. shot sampling 8* vs. *mean*) we can see quite the opposite result, *i.e.* the *shot sampling 8* approach produces the worst recommendation quality in comparison with other approaches while using a descriptor length of eight times larger than *mean* and *median*.

     These result indicates that by building a portability density function (PdF) of the feature values across shots, we are able to collect more useful information about video content in terms of their visual content in comparison with the basic approaches which aggregate the features by using the basic mean and median of features or by sampling a random number of shots where in the latter case, our results have shown that the excess of features can serve as a degrading factor.
   - **Explanation**: One of the main weak points of shot sampling approaches is that they collect information from a certain number of shots (*e.g.* 8 or 16 shots). Indeed, the final video feature descriptor components in this approach contains the information about certain shots, *not all*. This can be a limitation compared with PdF-based approaches which with the same descriptor length, contain information about *all* shots. The other limitation of shot-sampling based approaches is that, if we wish to increase to number of shots (*e.g.* 64, 128), we may end up not finding videos containing that number of shots. This is a limitating factor specially for movie trailers which usually have a short video duration (note that we need to sample the same number of shots from all videos in order to create a fixed-length descriptor for all videos).

**Table 4: Final length of different video descriptors.**

| Between-Shot Aggregation | Final Video Descriptor Length |
|---|---|
| mean | 6 |
| median | 6 |
| shot sampling 8 | 48 |
| shot sampling 16 | 96 |
| PdF 8 | 48 |
| PdF 16 | 96 |

## 3.3 Sensitivity to the Shot Segmentation Level

The second experiment comprises of studying the sensitivity of video recommendation quality with respect to the shot segmentation granularity. The segmentation level can hugely affect the number of shots and key frames produced after shot segmentation which in turn can affect the quality of recommendation. We segmented all videos into shots from a coarse to fine fashion. For movie trailers, coarse segmentation produces a small number of shots while fine segmentation can result in generation of a large number of shots. We show the recommendation quality as function of between-shot aggregation and five segmentation levels namely: *very coarse*, *coarse*, *normal*, *fine* and *very fine* by using 2D heat-map plot. As can be seen in Figure 3, there exists a significant difference in recommendation quality when moving from *very coarse* segmentation to *very fine* segmentation. In fact, while the *very coarse* and

*fine* and *very fine* segmentations produce low-quality recommendations, the *coarse* and *normal* segmentations produce the best results altogether.

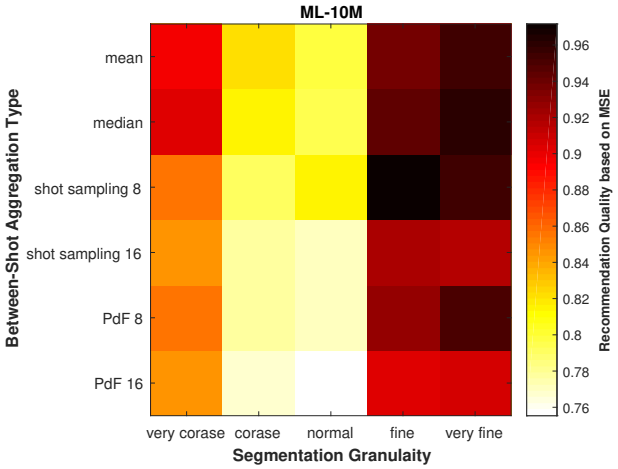


**Figure 3: Quality of recommendation *w.r.t.* (MSE) as function of the segmentation granularity level.**

## 4 CONCLUSION AND FUTURE WORK

In this paper, we studied the impact of using different video summarization models for the task of content-based video recommendation based on visual features. We presented preliminary results of an analysis involving applying different video summarization models in a content-based recommender system using a large existing real RS dataset (MovieLens-10M) and showed that recommendation quality can be significantly improved if we represent features across shots using a probability density function and choose a segmentation granularity which is not very fine nor very coarse. These results improve our understanding on the effect video summarization model on recommendation quality and provide a strong argument for exploring the potential of video summarization models in the design of video recommender systems based on visual features.

For future work, we plan to extend our work in a number of directions:

- We will widen our analysis by adopting bigger and different datasets (e.g. Netflix, Amazon), to provide a more robust statistical support to our finding.
- We will investigate the impact of using different video summarization models for full movies which contain significantly higher number of shots.
- We will extend the range of features extracted to include both visual and audio features, using features based on the MPEG-7 standard and DNN [6].
- We will analyze quality of recommender systems based on low-level features not only in terms of MSE, but also in terms of accuracy metrics (precision, recall, Map) in addition to novelty and diversity.

## ACKNOWLEDGMENTS

This work has been partially funded by Telecom Italia S.p.A., Services Innovation Department, Joint Open Lab S-Cube, Milano, and by the Austrian Science Fund (FWF): P25655.

## REFERENCES

[1] Bryan R Burnham, James H Neely, Yelena Naginsky, and Matthew Thomas. 2010. Stimulus-driven attentional capture by a static discontinuity between perceptual groups. *Journal of Experimental Psychology: Human Perception and Performance* 36, 2 (2010), 317.

[2] Manfred Del Fabro and Laszlo Böszörmenyi. 2013. State-of-the-art and future challenges in video scene detection: a survey. *Multimedia systems* 19, 5 (2013), 427–454.

[3] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, Pietro Piazzolla, and Massimo Quadrana. 2016. Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics* 5, 2 (2016), 99–113.

[4] Yashar Deldjoo, Mehdi Elahi, Massimo Quadrana, Paolo Cremonesi, and Franca Garzotto. 2015. Toward effective movie recommendations based on mise-en-scène film styles. In *Proceedings of the 11th Biannual Conference on Italian SIGCHI Chapter*. ACM, 162–165.

[5] Yashar Deldjoo, Fatemeh Nazary, and Ali M Fotouhi. 2015. A novel fuzzy-based smoke detection system using dynamic and static smoke features. In *Electrical Engineering (ICEE), 2015 23rd Iranian Conference on*. IEEE, 729–733.

[6] Yashar Deldjoo, Massimo Quadrana, Mehdi Elahi, and Paolo Cremonesi. 2017. Using Mise-En-Scene Visual Features based on MPEG7 and Deep Learning for Movie Recommendation. *arXiv preprint arXiv:1704.06109* (2017).

[7] Naveed Ejaz, Tayyab Bin Tariq, and Sung Wook Baik. 2012. Adaptive key frame extraction for video summarization using an aggregation mechanism. *Journal of Visual Communication and Image Representation* 23, 7 (2012), 1031–1040.

[8] P Geetha and Vasumathi Narayanan. 2008. A survey of content-based video retrieval. (2008).

[9] F Maxwell Harper and Joseph A Konstan. 2016. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4 (2016), 19.

[10] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. 2011. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41, 6 (2011), 797–819.

[11] Hansung Lee, Jaehak Yu, Younghee Im, Joon-Min Gil, and Daihee Park. 2011. A unified scheme of shot boundary detection and anchor shot detection in news video story parsing. *Multimedia Tools and Applications* 51, 3 (2011), 1127–1145.

[12] Arthur G Money and Harry Agius. 2008. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation* 19, 2 (2008), 121–143.

[13] István Pilászy, Dávid Zibriczky, and Domonkos Tikk. 2010. Fast als-based matrix factorization for explicit and implicit feedback datasets. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 71–78.

[14] Tim Pohle, Dominik Schnitzer, Markus Schedl, Peter Knees, and Gerhard Widmer. 2009. On Rhythm and General Music Similarity. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*. Kobe, Japan.

[15] Zeeshan Rasheed, Yaser Sheikh, and Mubarak Shah. 2005. On the use of computable features for film classification. *IEEE Transactions on Circuits and Systems for Video Technology* 15, 1 (2005), 52–64.

[16] Klaus Seyerlehner, Gerhard Widmer, Markus Schedl, and Peter Knees. 2010. Automatic Music Tag Classification based on Block-Level Features. In *Proceedings of the 7th Sound and Music Computing Conference (SMC)*. Barcelona, Spain.

[17] Jônatas Wehrmann, Rodrigo C Barros, Gabriel S Simões, Thomas S Paula, and Duncan D Ruiz. 2016. (Deep) Learning from Frames. In *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on*. IEEE, 1–6.

[18] Bo Yang, Tao Mei, Xian-Sheng Hua, Linjun Yang, Shi-Qiang Yang, and Mingjing Li. 2007. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 73–80.

[19] Howard Zhou, Tucker Hermans, Asmita V Karandikar, and James M Rehg. 2010. Movie genre classification via scene categorization. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 747–750.